Proceedings

THIRTY-FIFTH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL AND COMPUTING

September 29, September 30, and October 1, 1997

Allerton House, Monticello, Illinois
Sponsored by the
Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering of the
University of Illinois at Urbana-Champaign

PROCEEDINGS

THIRTY-FIFTH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL, AND COMPUTING

Tamer Başar
Alexander Vardy
Conference Co-Chairs

Conference held
September 29, September 30, and October 1, 1997
Allerton House
Monticello, Illinois

Sponsored by
The Coordinated Science Laboratory
and
The Department of Electrical and Computer Engineering
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign

621.3811
AL5
1997

FOREWORD

These Proceedings contain papers that were presented at the Thirty-fifth Annual Allerton Conference on Communication, Control, and Computing which was held at Allerton House, Monticello, Illinois, on September 29, September 30, and October 1, 1997. As in past years, the Conference was sponsored by the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign.

Sixty-six invited papers and sixty-six contributed papers were presented at the Conference in twenty-three sessions. We thank all the authors and particularly the speakers for contributing to the success of the Conference. We are also grateful to our colleagues at the University of Illinois at Urbana-Champaign: Professors A. Alleyne, B. Bamieh, T.Ü. Başar, R.E. Blahut, N. Boston, B. Hajek, I. Hajj, W.K. Jenkins, D. Jones, P.R. Kumar, U. Madhow, J.V. Medanic, D. Munson, P. Moulin, W.R. Perkins, D.V. Sarwate, N. Shanbhag, M.W. Spong, R.S. Sreenivas, R. Srikant, and P.G. Voulgaris, for organizing many of the invited sessions, reviewing the contributed papers, chairing sessions, and helping in many other ways in the smooth running of the Conference. Additional invited sessions were organized by Professor J. Walker of University of Nebraska, and Professors K. Park, W. Szpankowksi, J.V. Krogmeier, and S.B. Gelfand, all of Purdue University; we are especially grateful to them for their efforts on behalf of the Conference. Professors S. Verdú of Princeton University, M. Radenkovic of University of Colorado, M. Honig of Northwestern University, R.J. McEliece of the California Institute of Technology, and Dr. G.D. Forney of Motorola, Inc., chaired contributed and invited sessions at the Conference, and for this we thank them.

The plenary speaker at the Conference was Graham Goodwin of the University of Newcastle, Australia, who gave an illuminating lecture titled "Fundamental Limitations in Filtering and Control." We are grateful to him for his excellent presentation. We are also indebted to the string players from the "String Society," Urbana, for giving us a wonderful evening of classical music.

The Allerton Conference featured this year two special events. One was a two-day Workshop on "Future Directions in Systems and Control" organized by P.R. Kumar, and held on the Allerton grounds on September 28 and 29. The second day of this Workshop overlapped with the first day of the Allerton Conference. This gave the attendees the unique opportunity to also meet with and listen to some of the speakers at the Workshop. The second event was the dedication of a time capsule on the occasion of the 35th anniversary of the Allerton Conference. The time capsule was buried on the Allerton grounds at 6:00p.m. on Tuesday, September 30, to be reopened 35 years later, in the year 2032. Detailed information on the time-capsule dedication ceremonies and the burial can be found in the article on page *ii* of these Proceedings. The *Time Capsule Committee* was chaired by Bruce Hajek, to whom we are indebted for the superb organization of this unique event in the history of the Allerton Conference.

We thank Mrs. Beverly Becker and her staff at Allerton House for cheerfully accommodating our Conference and for their constant attention to our needs. It is also a great pleasure to thank all the staff members of the Coordinated Science Laboratory who so ably handled all the practical aspects of organizing the Conference. In particular, we wish to mention Mr. Steven Williams and Ms. Susan Baxley who handled all the business affairs, and Mr. Dan Jordan who handled the sales and shipping of these Proceedings. Most of all, however, we would like to thank our Conference Secretary, Mrs. Becky Lonberger, who with the able assistance of Mrs. Francie Bridges and Mrs. Nancy Carr worked all year round in handling announcements, calls for papers, program preparation, Allerton House arrangements, registration, and the final editing of these Proceedings. The lion's share of the credit for the success of the Conference belongs to these individuals and it gives us the greatest pleasure to acknowledge this here.

Before concluding this foreword, we are now launching here a new tradition for these Proceedings, to be carried from one year to the next: *Countdown for the reopening of the time capsule.* The number of years to go is 35. The countdown has started!

Tamer Başar and Alexander Vardy

ALLERTON CONFERENCE TIME-CAPSULE

On Tuesday, September 30, 1997, a time capsule was buried to celebrate thirty-five years of Allerton Conferences. The capsule is a stainless steel box, about one foot per side, with the engraving, "35th Annual Allerton Conference on Communication, Control and Computing, September 1997." It is buried three feet below the ground. The site is marked by a bronze plaque reading:

On the occasion of the 35th Annual Allerton Conference
on Communication, Control, and Computing, here is buried
a time capsule to be opened in thirty five years.

• Dedicated September 30, 1997 •

The plaque is set in a 200 pound concrete block. It is near ground level, but may become covered with a layer of dirt or grass. The capsule can be reached by taking 35 paces North from the Allerton House, along a line extending the western face of the House. It is on a gently sloping hill, ten feet north from the base of a lone tree. The GPS coordinates (WGS84) of the time capsule are[1] 40 degrees 00 minutes 15.39750 seconds North latitude, 88 degrees 38 minutes 41.74234 seconds West longitude.

The time capsule dedication ceremony began at 5p.m., with short introductory remarks by Prof. Bruce Hajek. Prof. Tamer Başar then commented on the early history of the Allerton Conference. The preface to the First Allerton Conference Proceedings is evidence to the concept and vision of Prof. Mac VanValkenburg, as founder of the Conference. Tables of contents of early conferences include the names of leading (and emerging) researchers, who attended those early meetings. Prof. Başar noted that the Allerton Conference started with 11 technical sessions and doubled this number within two years, thus reaching its current size of 22-25 sessions. He also presented a list of all Allerton Conference past chairs/co-chairs (Table 1). Prof. William R. Perkins, perhaps the only person to have attended all thirty-five Allerton Conferences, then noted that Mac VanValkenburg founded the Allerton Conference when there were very few conferences in the systems area of electrical engineering. The Allerton Conference thrived due to Mac's foresight and nurturing, and Prof. Perkins acknowledged our debt to Mac, who died this spring.

The next speaker in the dedication ceremony was Prof. Anthony Ephremides, who briefly summarized the state of technology and society 35 years ago. He pointed out that an international phone call was extremely expensive and required a waiting period for an operator to call back. Of course, the phone was rotary. A household radio receiver was a piece of furniture. The Berlin wall was up, and no one had yet set foot on the moon. There was general pessimism regarding the future, and most people believed that nuclear holocaust is likely to occur within the next 35 years. There was, however, an optimism about the future of technology.

Among the contents of the capsule are questionnaires filled out by conference participants, who were asked to make predictions about the future. Section I of the questionnaire is shown in Figure 1. Prof. Ephremides went on to answer the questions on this questionnaire, from the perspective of someone in 1963, the year of the first Conference. He pointed out that several items on the questionnaire would not have made any sense at all. For example, the notion of cloning animals seemed quite unrealistic at the time. There was no "information superhighway," no public-key cryptography, very little multiuser information theory and the state-space approach to controller design was not yet adopted.

Prof. Robert J. McEliece summarized the responses on the questionnaires for the audience. The tallys of the responses to Section I are also shown in Figure 1 (last minute submissions are not included in the tally). Prof. McEliece pointed out that most respondents thought that nuclear holocaust will not happen within the next 35 years. This was one of the two items that ranked highest in overall confidence of the respondents. The other item respondents were most sure about is that the state-space approach to controller design will not be abandoned within the next 35 years. Prof. McEliece summarized the many predictions offered by respondents in Sections II and III of the survey. The prevailing response to the question about the main difference between everyday life in 2032 and 1997 was that many respondents would be dead in 2032. The question concerned with how research will be funded in the year 2032 was answered this way in the response of Prof. Ephremides: "Thirty five years from now, researchers would pay 1% of their adjusted gross income as a special tax to finance the funding of research. In that way, only the truly committed researchers would continue performing research." Several responses to the question of how most individuals will be connected to the information superhighway involved various forms of mental telepathy.

---

[1] We are grateful to our colleagues from the Illinois Department of Natural Resources, and in particular to Dr. Christopher J. Stohr, for providing these GPS coordinates.

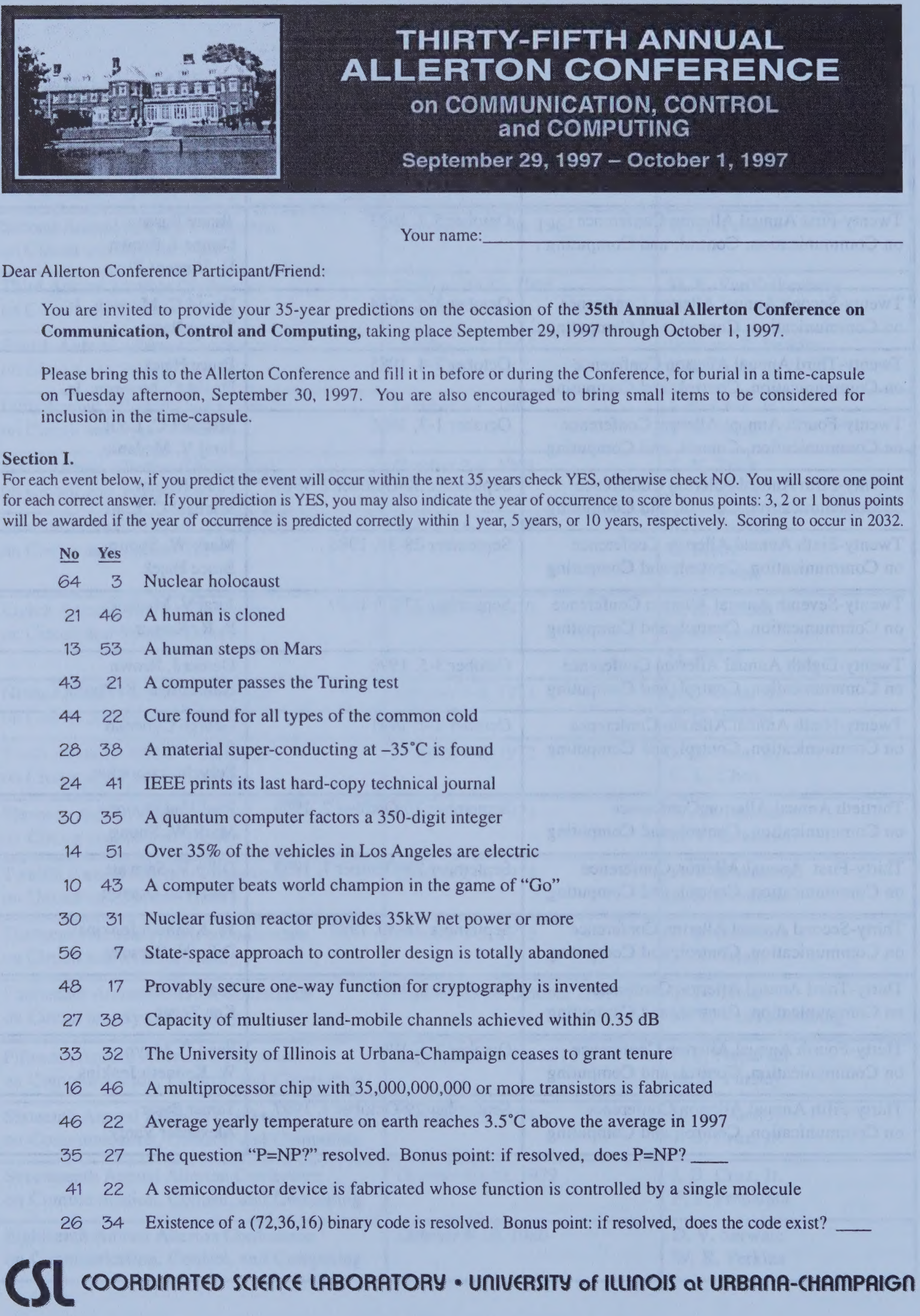
Finally, Prof. Richard E. Blahut went through the contents of the capsule and remarked on the significance of each item. The inside cover of the capsule has a University of Illinois inventory tag, since such administrative devices will surely be still needed in the future. A mock-up of a disk for two-dimensional flash optical storage, produced by Information Optics, was contributed by Prof. Blahut himself. A brochure describing an instructional robot marketed by the company of Prof. Mark Spong was included. Another brochure describes the Motorola Star-Tac phone as "the world's smallest cellular telephone." Several editions of popular magazines dedicated to web browsing, a CD, and a Texas Instruments overdrive chip were also included. Two items were placed in the capsule to be given to the person (or the estate of the person) with the highest scoring predictions. One was a bottle of port wine, with a recommendation that it be shared among those opening the capsule, and the other is a proof 1997 American Silver Eagle coin. In the last minute or two, several other items were added by members of the audience. These include a dress tie (perhaps because such things will not be worn in 35 years), a driver's license and a credit card (from two different people), a blank check from a local bank, and a receipt for Corky's Limousine Service, which over the years has shuttled many participants to the Allerton Conference.

The time-capsule committee consisted of:
Tamer Başar
Bruce Hajek (chair)
David Munson
Alexander Vardy

Table 1: 35 YEARS OF ALLERTON CONFERENCES

Name	Dates	Chair(s)
First Annual Allerton Conference on Circuit and System Theory	November 15-17, 1963	J. B. Cruz, Jr.
Second Annual Allerton Conference on Circuit and System Theory	September 28-30, 1964	W. R. Perkins
Third Annual Allerton Conference on Circuit and System Theory	October 20-22, 1965	M. E. VanValkenburg
Fourth Annual Allerton Conference on Circuit and System Theory	October 5-7, 1966	William R. Perkins J. B. Cruz, Jr.
Fifth Annual Allerton Conference on Circuit and System Theory	October 4-6, 1967	J. B. Cruz, Jr. T. N. Trick
Sixth Annual Allerton Conference on Circuit and System Theory	October 2-4, 1968	T. N. Trick R. T. Chien
Seventh Annual Allerton Conference on Circuit and System Theory	October 8-10, 1969	A. H. Haddad G. Metze T. N. Trick
Eighth Annual Allerton Conference on Circuit and System Theory	October 7-9, 1970	A. H. Haddad B. C. Kuo G. Metze
Ninth Annual Allerton Conference on Circuit and System Theory	October 6-8, 1971	J. B. Cruz, Jr. A. H. Haddad
Tenth Annual Allerton Conference on Circuit and System Theory	October 4-6, 1972	J. B. Cruz, Jr. C. L. Chen
Eleventh Annual Allerton Conference on Circuit and System Theory	October 3-5, 1973	W. R. Perkins P. V. Kokotovic
Twelfth Annual Allerton Conference on Circuit and System Theory	October 2-4, 1974	P. V. Kokotovic E. S. Davidson
Thirteenth Annual Allerton Conference on Circuit and System Theory	October 1-3, 1975	E. S. Davidson T. N. Trick
Fourteenth Annual Allerton Conference on Circuit and System Theory	September 29-October 1, 1976	T. N. Trick M. E. VanValkenburg
Fifteenth Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1977	F. P. Preparata M. B. Pursley
Sixteenth Annual Allerton Conference on Communication, Control, and Computing	October 4-6, 1978	M. B. Pursley J. B. Cruz, Jr.
Seventeenth Annual Allerton Conference on Communication, Control, and Computing	October 10-12, 1979	J. B. Cruz, Jr. F. P. Preparata
Eighteenth Annual Allerton Conference on Communication, Control, and Computing	October 8-10, 1980	D. V. Sarwate W. R. Perkins

Nineteenth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1981	W. K. Jenkins D. V. Sarwate
Twentieth Annual Allerton Conference on Communication, Control, and Computing	October 6-8, 1982	H. V. Poor W. K. Jenkins
Twenty-First Annual Allerton Conference on Communication, Control, and Computing	October 5-7, 1983	Tamer Başar Donna J. Brown H. Vincent Poor
Twenty-Second Annual Allerton Conference on Communication, Control, and Computing	October 3-5, 1984	David C. Munson, Jr. Tamer Başar
Twenty-Third Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1985	Bruce Hajek David C. Munson, Jr.
Twenty-Fourth Annual Allerton Conference on Communication, Control, and Computing	October 1-3, 1986	Michael C. Loui Juraj V. Medanic
Twenty-Fifth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1987	P. R. Kumar Michael C. Loui
Twenty-Sixth Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1988	Mark W. Spong Bruce Hajek
Twenty-Seventh Annual Allerton Conference on Communication, Control, and Computing	September 27-29, 1989	Juraj V. Medanic P. R. Kumar
Twenty-Eighth Annual Allerton Conference on Communication, Control, and Computing	October 3-5, 1990	Donna J. Brown Kameshwar R. Poolla
Twenty-Ninth Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1991	George Cybenko Constantine Polychronopoulos
Thirtieth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1992	Paul Van Dooren Mark W. Spong
Thirty-First Annual Allerton Conference on Communication, Control, and Computing	September 29-October 1, 1993	Dilip V. Sarwate Paul Van Dooren
Thirty-Second Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1994	W. Kenneth Jenkins Dilip V. Sarwate
Thirty-Third Annual Allerton Conference on Communication, Control, and Computing	October 4-6, 1995	Sean Meyn Ken Zeger
Thirty-Fourth Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1996	Sean P. Meyn W. Kenneth Jenkins
Thirty-Fifth Annual Allerton Conference on Communication, Control, and Computing	September 29-October 1, 1997	Tamer Başar Alexander Vardy



THIRTY-FIFTH ANNUAL
ALLERTON CONFERENCE
on COMMUNICATION, CONTROL
and COMPUTING
September 29, 1997 – October 1, 1997

Your name:______________________

Dear Allerton Conference Participant/Friend:

You are invited to provide your 35-year predictions on the occasion of the **35th Annual Allerton Conference on Communication, Control and Computing,** taking place September 29, 1997 through October 1, 1997.

Please bring this to the Allerton Conference and fill it in before or during the Conference, for burial in a time-capsule on Tuesday afternoon, September 30, 1997. You are also encouraged to bring small items to be considered for inclusion in the time-capsule.

Section I.

For each event below, if you predict the event will occur within the next 35 years check YES, otherwise check NO. You will score one point for each correct answer. If your prediction is YES, you may also indicate the year of occurrence to score bonus points: 3, 2 or 1 bonus points will be awarded if the year of occurrence is predicted correctly within 1 year, 5 years, or 10 years, respectively. Scoring to occur in 2032.

No	Yes	
64	3	Nuclear holocaust
21	46	A human is cloned
13	53	A human steps on Mars
43	21	A computer passes the Turing test
44	22	Cure found for all types of the common cold
28	38	A material super-conducting at –35°C is found
24	41	IEEE prints its last hard-copy technical journal
30	35	A quantum computer factors a 350-digit integer
14	51	Over 35% of the vehicles in Los Angeles are electric
10	43	A computer beats world champion in the game of "Go"
30	31	Nuclear fusion reactor provides 35kW net power or more
56	7	State-space approach to controller design is totally abandoned
48	17	Provably secure one-way function for cryptography is invented
27	38	Capacity of multiuser land-mobile channels achieved within 0.35 dB
33	32	The University of Illinois at Urbana-Champaign ceases to grant tenure
16	46	A multiprocessor chip with 35,000,000,000 or more transistors is fabricated
46	22	Average yearly temperature on earth reaches 3.5°C above the average in 1997
35	27	The question "P=NP?" resolved. Bonus point: if resolved, does P=NP? ____
41	22	A semiconductor device is fabricated whose function is controlled by a single molecule
26	34	Existence of a (72,36,16) binary code is resolved. Bonus point: if resolved, does the code exist? ____

CSL COORDINATED SCIENCE LABORATORY • UNIVERSITY of ILLINOIS at URBANA-CHAMPAIGN

Figure 1

TABLE OF CONTENTS

ACKNOWLEDGEMENTS......i

ALLERTON CONFERENCE TIME-CAPSULE......i i

I-B: ALGEBRAIC GEOMETRY AND CODES
Organizers/Chairs: N. Boston/J. Walker
(University of Illinois/University of Nebraska)

EXPONENTIAL SUMS OVER FINITE RINGS: A REPORT......1
C.J. Moreno

EXPONENTIAL SUM ESTIMATES AND APPLICATIONS......14
W.-C.W. Li

EXPLICIT MODULAR TOWERS......23
N.D. Elkies

ON THE MISSING FUNCTIONS OF A PYRAMID OF CURVES......33
R. Pellikaan

QUADRATIC RESIDUE CODES AND HYPERELLIPTIC CURVES......41
I.M. Duursma

ON THE DERMINATION OF ERROR VALUES FOR CODES FROM A CLASS OF MAXIMAL CURVES......44
R. Koetter

I-C: WIRELESS COMMUNICATIONS: ADAPTIVE RECEPTION TECHNIQUES
Organizers: U. Madhow/D. Jones
(University of Illinois at Urbana-Champaign)
Chair: D. Jones
(University of Illinois at Urbana-Champaign)

COMPUTATIONALLY EFFICIENT BLIND EQUALIZATION......54
D.R. Brown, P.B. Schniter, and C.R. Johnson, Jr.

EQUALIZATIONS IN WIRELESS ATM......64
L. Tong and J.Q. Bao

COUPLED ADAPTATION IN WIRELESS MULTIPLE ACCESS SYSTEMS......74
G.J. Pottie

BLIND ADAPTIVE SPACE-TIME LINEAR MULTIUSER DEMODULATION......83
M.L. Honig, R. Ratasuk, and S. Verdú

JOINT TRANSMITTER AND RECEIVER BEAMFORMING FOR MAXIMUM CAPACITY IN SPATIAL DIVISION MULTIACCESS......93
J.-H. Chang, L. Tassiulas, and F. Rashid-Farrokhi

I-D: STATISTICAL SIGNAL PROCESSING

Chair: P. Moulin
(University of Illinois at Urbana-Champaign)

QUANTIZATION FOR DISTRIBUTED ESTIMATION WITH COMMUNICATION AND STORAGE CONSTRAINTS 102
V. Megalooikonomou and Y. Yesha

ORIENTATION ESTIMATION BOUNDS FOR HIGH RESOLUTION RADAR DATA 112
J.A. O'Sullivan, V. Kedia, and S.P. Jacobs

SOURCE SEPARATION OF SPEECH SIGNALS USING KURTOSIS MAXIMIZATION 122
J.P. LeBlanc and P.L. De Leòn

DISCRETE SYMBOL QUANTITY AND THE MINIMUM PROBABILITY OF ERROR FOR A COMBINED INFORMATION CLASSIFICATION TEST 124
R. Lynch and P. Willett

JOINT LOSSLESS-SOURCE AND CHANNEL CODING USING ARITHMETIC CODING 126
G.F. Elmasry

COEFFICIENT RATE IN TRANSFORM CODING 128
W. Yang and J.D. Gibson

MODELING VIDEO TRAFFIC USING WAVELETS 138
S. Ma and C. Ji

I-E: CODING THEORY

Chair: B. Frey
(University of Illinois at Urbana-Champaign)

OPTIMAL GENERATOR MATRICES OF MULTI-BIT INFORMATION BLOCKS IN THE GOLAY CODE 147
H. Chen and J.T. Coffey

BINOMIAL BOUNDS ON SPECTRA OF CODES WITH KNOWN DUAL DISTANCE 156
I. Krasikov and S. Litsyn

SEMIGROUP RINGS AND THE EXTENSION THEOREM FOR LINEAR CODES 163
J.A. Wood

ON THE ALGEBRAIC STRUCTURE OF 2-OBSERVABLE CONTROLLABLE GROUP CODES 173
S. Calabrò, M. Campanella, and G. Garbo

MULTIDIMENSIONAL CONVOLUTIONAL CODES 183
P.A. Weiner

RECURSIVE MAXIMUM LIKELIHOOD DECODING ALGORITHM FOR CONVOLUTIONAL CODES AND ITS COMPLEXITY 185
H. Yamamoto, T. Fujiwara, and T. Uchida

NUMERICAL ANALYSIS OF THE MAXIMUM A POSTERIOR ALGORITHM 187
M.C. Reed, P.D. Alexander, and J.A. Asenstorfer

II-B: ALGEBRAIC-GEOMETRIC CODES AND DECODING ALGORITHMS

Organizers: R.E. Blahut/A. Vardy
(University of Illinois at Urbana-Champaign)
Chair: R.E. Blahut
(University of Illinois at Urbana-Champaign)

CURVES WITH MANY POINTS.......196
I.F. Blake

THE APPLICATIONS OF GENERALIZED BEZOUT'S THEOREM TO THE CODES FROM THE CURVES IN HIGH DIMENSIONAL SPACES.......205
G.-L. Feng, J. Zhu, X. Shi, and T.R.N. Rao

DECODING REED SOLOMON CODES BEYOND THE ERROR-CORRECTION DIAMETER.......215
M. Sudan

DECODING ALGEBRAIC-GEOMETRIC CODES BEYOND THE ERROR-CORRECTION BOUND.......225
M.A. Shokrollahi and H. Wasserman

FAST ERASURE AND ERROR DECODING OF EVALUATION CODES.......229
T. Høholdt

FAST GMD DECODING OF ALGEBRAIC CODES.......231
S. Sakata

II-C: SELF-SIMILAR TRAFFIC MODELS

Organizers: K. Park/W. Szpankowski
(Purdue University)
Chair: K. Park
(Purdue University)

QUEUEING ANALYSIS WITH TRAFFIC MODELS BASED ON DETERMINISTIC DYNAMICAL SYSTEMS.......233
V. Anantharam

ANALYTIC INFORMATION THEORY IN SERVICE OF QUEUEING WITH AGGREGATED EXPONENTIAL ON/OFF ARRIVALS.......242
P. Jacquet

FRACTAL NETWORK TRAFFIC MODELING: PAST, PRESENT, AND FUTURE.......252
B. Ryu

AN EMPIRICAL INTERNET TRAFFIC STUDY.......261
J. Aracil, R. Edell, and P. Varaiya

LOOKING BEHIND AND BEYOND SELF-SIMILARITY: ON SCALING PHENOMENA IN MEASURED WAN TRAFFIC.......269
A. Feldmann, A.C. Gilbert, W. Willinger, and T.G. Kurtz

II-D: MULTIUSER DETECTION

Chair: S. Verdú
(Princeton University)

MULTIUSER DEMODULATION: EFFECTIVE INTERFERENCE, EFFECTIVE BANDWIDTH AND CAPACITY.......281
D. Tse and S. Hanly

IDENTIFIABILITY PROBLEM IN BLIND SEPARATION OF SYNCHRONOUS BPSK SIGNALS IN DIGITAL WIRELESS COMMUNICATION 291
S. Kwon and D.R. Fuhrmann

A RECURSIVE DECORRELATING DETECTOR FOR AN ASYNCHRONOUS MULTIUSER COMMUNICATION SYSTEM 301
S.V. Uppala and J.D. Sahr

MULTIUSER DECISION-FEEDBACK RECEIVERS FOR ASYNCHRONOUS CDMA SYSTEMS OVER MISMATCHED RAYLEIGH FADING CHANNEL 311
T.-T. Chen and J.S. Lehnert

REDUCED COMPLEXITY MAXIMUM LIKELIHOOD RECEIVER FOR CODED MULTIUSER COMMUNICATION SYSTEMS 321
M.-H. Yang and M. Brandt-Pearce

FURTHER ANALYSIS OF MUSIC-BASED DELAY ESTIMATORS FOR DS-CDMA SYSTEMS 330
L.-C. Chu and U. Mitra

II-E: COMMUNICATION SYSTEMS
Chair: D. Sarwate
(University of Illinois at Urbana-Champaign)

A NOVEL RECEIVER FOR BURST MODE DATA DETECTION IN THE PRESENCE OF TIMING OFFSET 340
C.R. Nassar and M.R. Soleymani

SIGNAL DESIGN FOR IMPROVED SYNCHRONIZATION IN FREQUENCY HOPPED SPREAD SPECTRUM CDMA COMMUNICATIONS 350
L.A. Osadciw and E.L. Titlebaum

OPTIMIZING INFORMATION EFFICIENCY IN A MOBILE PACKET RADIO NETWORK 360
M.W. Subbarao and B.L. Hughes

ROBUST ADAPTIVE CODED MODULATION FOR TIME-VARYING CHANNELS WITH DELAYED FEEDBACK 370
D.L. Goeckel

ADAPTIVE CHANNEL CODING IN DS-CDMA NETWORKS UNDER IMPERFECT STATE INFORMATION 380
S.-J. Oh and K.M. Wasserman

OPTIMAL POWER ADAPTATION IN DIGITAL REPEATERS 383
D.L. Goeckel

REDUNDANT FILTERBANK PRECODERS AND EQUALIZERS: OPTIMAL DESIGNS AND BLIND SIGNAL RECOVERY 385
A. Scaglione, G.B. Giannakis, and S. Barbaross

III-A: COMBINATORIAL AND ALGEBRAIC STRUCTURE OF CODES
Organizers: R.E. Blahut/A. Vardy
(University of Illinois at Urbana-Champaign)
Chair: A. Vardy
(University of Illinois at Urbana-Champaign)

ON INFINITE FAMILIES OF 3-DESIGNS FROM PREPARATA CODES AND RELATED CODES OVER Z_4 395
T. Helleseth, P.V. Kumar, C. Rong, and K. Yang

THE DECOMPOSITION OF Z_4 CODES 404
C.W. Huffman

DISTANCE REGULAR SPHERICAL CODES 413
T. Ericson

SPLIT WEIGHT ENUMERATORS OF EXTREMAL SELF-DUAL CODES 422
J. Fields and V. Pless

ON CODES FOR ERROR CORRECTION AND BLOCK SYNCHRONIZATION 432
Ø. Ytrehus

COMBINED ARRAY SIGNAL PROCESSING AND SPACE-TIME CODING 440
V. Tarokh, A. Naguib, N. Seshadri, and A.R. Calderbank

III-B: WIRELESS COMMUNICATIONS: PROBLEMS IN DETECTION AND ESTIMATION

Organizers: U. Madhow/D. Jones
(University of Illinois at Urbana-Champaign)
Chair: U. Madhow
(University of Illinois at Urbana-Champaign)

ADAPTIVE MULTIUSER DETECTION IN NON-GAUSSIAN CHANNELS 450
X. Wang and H.V. Poor

AN OPTIMAL ONE-SHOT MULTIUSER DETECTOR FOR FLAT RAYLEIGH FADING CHANNELS 460
A. Sendonaris and B. Aazhang

MULTIUSER DETECTION WITH RANDOM SPREADING AND ERROR-CORRECTION CODES: FUNDAMENTAL LIMITS 470
S. Verdú and S. Shamai

DISTRIBUTED ESTIMATION TECHNIQUES FOR WIRELESS SENSOR NETWORKS 483
H.C. Papadopoulos and G.W. Wornell

DECISION FEEDBACK SEQUENCE ESTIMATION FOR UNWHITENED ISI CHANNELS 493
A. Hafeez and W.E. Stark

BLIND CHANNEL IDENTIFICATION FOR NARROWBAND DIGITAL COMMUNICATIONS BASED ON PARAMETRIC MODELING OF THE CHANNEL IMPULSE RESPONSE 503
M.D. Zoltowski and D.-F. Tseng

III-C: DISCRETE EVENT DYNAMICAL SYSTEMS

Organizer/Chair: R.S. Sreenivas
(University of Illinois at Urbana-Champaign)

SUPERVISORY CONTROL OF CONCURRENT TASK SEQUENCES OPERATING OVER LOW-LEVEL DEVICES 513
J. Ashley Jr. and L.E. Holloway

MULTIPLE MARKED LANGUAGES AND NONINTERFERENCE IN MODULAR SUPERVISORY CONTROL 523
J.G. Thistle and R.P. Malhamé

A SUPERVISORY CONTROL SYNTHESIS CASE STUDY: THE ANTENNA CONTROL SYSTEM 533
M. Barbeau, M. Frappier, F. Kabanza, and R. St-Denis

ON SUPERVISORY CONTROL OF PARTIAL SPECIFICATION ARISING IN PROTOCOL CONVERSION 543
R. Kumar and M. Fabian

PERFORMANCE OPPORTUNITY FOR FLEXIBLE WORKERS 553
M.P. Van Oyen, E. Senturk-Gel, and W.J. Hopp

III-D: COMMUNICATION NETWORK ALGORITHMS
Chair: B. Hajek
(University of Illinois at Urbana-Champaign)

SCHEMES FOR RELIABLE FAULT DETECTION IN COMMUNICATION NETWORKS 563
M.-S. Oh and P.S. Min

ON APPROXIMATION ALGORITHMS FOR RADIO NETWORK SCHEDULING 573
A. Sen and E. Malesinska

EFFECT OF PACKET LENGTHS AND POWER LEVELS ON RANDOM MULTIPLE ACCESS 583
W. Luo and A. Ephremides

EFFICIENT SORTING ON THE STAR GRAPH INTERCONNECTION NETWORK 585
S.G. Akl and T. Wolff

MATCHING OUTPUT QUEUEING WITH COMBINED INPUT AND OUTPUT QUEUEING 595
N. McKeown, B. Prabhakar, and M. Zhu

OPTIMAL SPACE PRIORITY POLICIES FOR SHARED MEMORY ATM SYSTEMS 604
R. Roy and S.S. Panwar

III-E: STABILITY
Chair: P. Voulgaris
(University of Illinois at Urbana-Champaign)

NECESSARY AND SUFFICIENT CONDITIONS FOR A SEMI-INTERVAL POLYTOPE, WITH ODD PART FIXED, TO CONTAIN A HURWITZ POLYNOMIAL 614
L.R. Pujara

A SUFFICIENT CONDITION FOR THE STABILITY OF MULTI-INPUT MULTI-OUTPUT ARMA MODELS WITH TIME-VARYING COEFFICIENTS 623
P.R. Bouthellier and B.K. Ghosh

DISCRETIZATION SCHEMES OF LYAPUNOV FUNCTIONAL IN THE STABILITY PROBLEM OF LINEAR UNCERTAIN TIME-DELAY SYSTEMS 631
K. Gu

CONTROL OF DISSYNCHRONIZED SYSTEMS OF DIFFERENCE EQUATIONS WITH DELAY 641
A.S.C. Sinha

SUFFICIENT CONDITIONS FOR THE EXPONENTIAL STABILITY OF LINEAR SYSTEMS SUBJECT TO NORM BOUNDED, BOUNDED RATE UNCERTAINTIES 643
F. Amato

STABILITY ROBUSTNESS OF SAMPLED-DATA SYSTEMS WITH PERIODIC-DIGITAL ELEMENTS 646
D.S. Ortiz and J.S. Freudenberg

A NEW APPROACH TO THE SYNTHESIS OF PARAMETER DEPENDENT LYAPUNOV FUNCTIONS FOR THE ESTIMATION OF THE STABILITY MARGIN IN THE PRESENCE OF SLOWLY-VARYING PARAMETERS 656
F. Amato and G. Celentano

IV-A: ITERATIVE DECODING AND TURBO CODES

Organizers: R.E. Blahut/A. Vardy
(University of Illinois at Urbana-Champaign)
Chairs: G.D. Forney, Jr./A. Vardy
(Motorola, Inc./University of Illinois)

FACTOR GRAPHS AND ALGORITHMS........666
B.J. Frey, F.R. Kschischang, H.-A. Loeliger, and N. Wiberg

LINEAR-TIME BINARY CODES CORRECTING LOCALIZED ERASURES........681
A. Barg and S. Zhou

THE CONVOLUTED HISTORY OF MINIMAL ENCODERS........687
H.-A. Loeliger

TRELLIS-CONSTRAINED CODES........691
B.J. Frey and D.J.C. MacKay

APPLICATION OF DISTANCE SPECTRUM ANALYSIS TO TURBO CODE PERFORMANCE IMPROVEMENT........701
M. Öberg and P.H. Siegel

ON DETERMINISTIC LINEAR INTERLEAVERS FOR TURBO-CODES........711
O.Y. Takeshita and D.J. Costello, Jr.

IV-B: WIRELESS COMMUNICATIONS: SYSTEM DESIGN AND OPTIMIZATION

Organizers: U. Madhow/D. Jones
(University of Illinois at Urbana-Champaign)
Chair: M. Honig
(Northwestern University)

AN ADAPTIVE TRANSMISSION METHOD FOR FREQUENCY-HOP RADIO NETWORKS........713
M.B. Pursley and C.S. Wilkins

SYNCHRONOUS COHERENT OFDM FOR BROADBAND CELLULAR RADIO SYSTEMS........719
K.L. Baum

DESIGNING USER CODES FOR MUI ELIMINATION AND UNKNOWN MULTIPATH MITIGATION IN ASYNCHRONOUS CDMA SYSTEMS........729
G.B. Giannakis and A. Scaglione

POWER CONTROL FOR AN ASYNCHRONOUS MULTI-RATE DECORRELATOR........739
M. Saquib, R. Yates, and A. Ganti

THE ROLE OF ERROR CORRELATIONS IN THE DESIGN OF PROTOCOLS FOR PACKET SWITCHED SERVICES........749
M. Zorzi and R.R. Rao

TCP IMPROVEMENTS FOR HETEROGENEOUS NETWORKS: THE DAEDALUS APPROACH........759
H. Balakrishnan, V.N. Padmanabhan, S. Seshan, M. Stemm, E. Amir, and R.H. Katz

IV-C: ADAPTIVE SIGNAL PROCESSING, IDENTIFICATION, AND CONTROL: COMMON PROBLEMS AND DIRECTIONS

Organizers: J.V. Krogmeier/S.B. Gelfand
(Purdue University)
Chair: J.V. Krogmeier
(Purdue University)

HARMONIC BALANCE METHODS FOR NONLINEAR FEEDBACK IDENTIFICATION 769
R.A. Casas, C.A. Jacobson, G.J. Rey, M. Green, and C.R. Johnson, Jr.

ALMOST-SURE PARAMETER ESTIMATION FOR PARTIALLY OBSERVED DISCRETE-TIME LINEAR SYSTEMS 779
R. Elliott, J. Ford, and J.B. Moore

WEIGHT REOPTIMIZATION IN OBE ALGORITHMS: AN INITIAL STUDY 789
D. Joachim, J.R. Deller, Jr., and M. Nayeri

AN ANALYSIS OF δ-STEP AHEAD ADAPTIVE PREDICTION AND ITS APPLICATION IN ACTIVE NOISE CONTROL 798
X. Kong and S.M. Kuo

NEW RESULTS ON STABILITY OF VARIABLE STEPSIZE LMS ALGORITHMS 807
S.B. Gelfand, J.V. Krogmeier, and Y. Wei

IV-D: CONTROL APPLICATIONS

Chair: A. Alleyne
(University of Illinois at Urbana-Champaign)

THE INFLUENCE OF FRICTION ON TIME OPTIMAL ROBOT TRAJECTORIES 817
B. Hartel, M. Steinbach, H.G. Bock, and R.W. Longman

A DESIGN OF C* ADAPTIVE FLIGHT CONTROL SYSTEM WITH UNCERTAIN DISTURBANCE 827
Y. Osa, S. Uchikado, K. Kanai, and K. Tanaka

APPLICATION OF H_∞ CONTROL AND CLOSED LOOP IDENTIFICATION TO A MULTIVARIABLE MAGNETIC LEVITATION SYSTEM 835
K. Nakashima, T. Tsujino, and T. Fujii

IMPLEMENTING TIME OPTIMAL ROBOT MANEUVERS USING REALISTIC ACTUATOR CONSTRAINTS AND LEARNING CONTROL 845
J. Li and R.W. Longman, V.H. Schultz, and H.G. Bock

IV-E: ROBUSTNESS AND PERFORMANCE LIMITATIONS

Chair: M. Radenković
(University of Colorado)

FUNDAMENTAL LIMITATIONS OF DISTURBANCE REJECTION IN FEEDBACK CONTROL SYSTEMS 848
D.E. Davison, P.T. Kabamba, and S.M. Meerkov

LIMITING PERFORMANCE OF OPTIMAL LINEAR FILTERS 858
J.H. Braslavsky, M.M. Seron, D.Q. Mayne, and P.V. Kokotović

A DESIGN TRADEOFF FOR SINGLE INPUT TWO OUTPUT FEEDBACK SYSTEMS 868
A.R. Woodyatt, J.S. Freudenberg, and R.H. Middleton

TOPOLOGICAL CONCEPTS FOR ROBUSTNESS MARGIN PROBLEMS 878
N.-P. Ke

SOLUTION OF THE STATE FEEDBACK SINGULAR H^{∞} CONTROL PROBLEM FOR LINEAR TIME-VARYING SYSTEMS 888
F. Amato, M. Mattei, and A. Pironti

PLENARY SESSION

FUNDAMENTAL DESIGN CONSTRAINTS IN LINEAR FILTERING AND CONTROL 898
G.C. Goodwin

V-A: STOCHASTIC SYSTEMS AND CONTROL
Chair: C. Humes
(Universidade de Sao Paulo)

DIFFERENTIAL TRAINING OF ROLLOUT POLICIES 913
D.P. Bertsekas

ESTIMATION OF STOCHASTIC-PARAMETER SIGNALS USING LINEAR MATRIX INEQUALITIES 923
J. Mohseni, E.E. Yaz, and K.J. Olejniczak

CONTROL ACCESS TO PERFORMANCE MEASURE STATISTICS: A SECOND MOMENT CASE 933
S.R. Liberty, P. Quast, and M.K. Sain

RISK-SENSITIVE QUEUEING 943
S.P. Coraluppi and S.I. Marcus

V-B: DECODING ALGORITHMS AND CONVOLUTIONAL CODES
Organizers: R.E. Blahut/A. Vardy
(University of Illinois at Urbana-Champaign)
Chair: R.J. McEliece
(California Institute of Technology)

CONSTRUCTION OF CONVOLUTIONAL CODES USING METHODS FROM LINEAR SYSTEMS THEORY 953
J. Rosenthal and R. Smarandache

MINIMUM WEIGHTS AND CORRECTABLE ERROR PATTERNS OF CONVOLUTIONAL CODES 961
J. Justesen

MULTILEVEL CODING: USE OF HARD DECISIONS IN MULTISTAGE DECODING 966
U. Wachsmann, R. Fischer, and J. Huber

ELLIPSOIDS IN HAMMING SPACES: ISOPERIMETRY, RANDOM SEARCH, AND MAXIMUM LIKELIHOOD DECODING 976
I. Dumer

V-C: SIGNAL PROCESSING AND COMMUNICATIONS FOR VLSI
Organizers: N. Shanbhag/I. Hajj
(University of Illinois at Urbana-Champaign)
Chair: N. Shanbhag
(University of Illinois at Urbana-Champaign)

LOW-POWER PULSE-SHAPING DIGITAL FILTERS 986
A. Willson, Jr.

HETEROGENEOUS DIGIT-SERIAL FINITE FIELD MULTIPLIERS AND LOW-ENERGY REED-SOLOMON CODECS 993
L. Song, K.K. Parhi, I. Kuroda, and T. Nishitani

REDUCED COMPLEXITY DECISION FEEDBACK EQUALIZATION FOR HIGH-SPEED COMMUNICATIONS OVER SPARSE CHANNELS 1003
I.J. Fevrier, S.B. Gelfand, and M.P. Fitz

V-D: WIRELESS CHANNEL MODELS
Chair: E. Biglieri
(Politecnico di Torino, Italy)

THE ROLE OF MEMORY IN ENTROPY APPROXIMATION FOR HIDDEN MARKOV MODELS 1013
B.M. Hochwald and P.R. Jelenković

POWER CONTROL FOR THE FADING CHANNEL: AN INFORMATION-THEORETICAL APPROACH 1023
G. Caire, G. Taricco, and E. Biglieri

MULTIPLE-ANTENNA COMMUNICATIONS WHEN NOBODY KNOWS THE RAYLEIGH FADING COEFFICIENTS 1033
T.L. Marzetta and B.M. Hochwald

CAPACITY OF CORRELATED JAMMING CHANNELS 1043
M. Médard

V-E: NONLINEAR SYSTEMS AND APPROXIMATION
Chair: M. Spong
(University of Illinois at Urbana-Champaign)

VOLTERRA FEEDBACK SYNTHESIS: A SYSTEMATIC ALGORITHM FOR SIMPLIFIED VOLTERRA CONTROLLER DESIGN AND REALIZATION 1053
P.M. Sain, M.K. Sain, and B.F. Spencer, Jr.

GENERALIZED ADJOINT SYSTEMS INVERSES 1063
D. Serakos

ON THE NEED FOR STABILIZING AND THE METHODS FOR STABILIZING REPETITIVE AND LEARNING CONTROLLERS 1067
R.W. Longman and Y.-P. Hsin

A SEPARATION PROPERTY OF OBSERVER/CONTROLLER FOR CONTINUOUS-TIME FUZZY SYSTEMS 1071
A. Jadbabaie, A. Titli, and M. Jamshidi

ON THE LOCAL POLYNOMIAL APPROXIMATION OF THE NORMAL FORM: THE SISO CASE 1079
J.A. Esquivel and J. Collado

LIST OF AUTHORS 1081

EXPONENTIAL SUMS OVER FINITE RINGS:

A REPORT

by Carlos Julio Moreno (CUNY)

1 Introduction

In this report we review the general theory of exponential sums over finite rings. We hope that our exposition will facilitate their applications. At any rate, we strongly believe that some of the ideas we describe deserve to be more widely known because of their intrinsic mathematical value. Two main themes motivate the contents and significance of our exposition and these will be briefly described below.

The first theme deals with an idea which appears explicitly in a beautiful expository article written by Richard Blahut in 1985 (see [1]) in which he expressed very clearly the need to approach the study of **digital signal processing, error control codes**, and **discrete Fourier transforms** without prejudicing the choice of field over which the computations are to be carried out. His insightful observations into the linear algebra structures that underline such constructions as **convolutional transforms,** and **Reed-Solomon codes**, just to mention a few, could be viewed as a proposal for an abstract theory of such objects which is free of fields of definition [1]. In the spirit of modern algebra, had Blahut suggested the replacement of fields by rings, he would have been close to fortelling some of the many developments in the last ten years concerning the use of finite rings in the construction of error control codes, digital signal processing, and finite Fourier transforms. We view some of the interesting recent work on error control codes over finite rings by Hammons, Kummar, Calderbank, Sloane, Sole, and Walker, as well as related applications found by W. Li and Oscar Moreno and others, as significant progress in the use of finite rings in accord with the directions implicitly suggested in Blahut's paper.

As for the second theme, we like to recall that the transmission of electromagnetic signals is based on Maxwell's theory of the wave equation, and in this theory, the **method of stationary phase** is used to extract from the solutions of the wave equation asymptotic information about the radiation emitted by a source when the frequency is large. This is a classical problem in the theory of partial differential equations and consists in analysing for large k the integrals of the type $\int_S \frac{e^{ik|x-y|}}{|x-y|}c(y)dA$, where S is the surface of the radiating body, dA is the surface area, and $c(y)$ is a function that contains information about the amplitude and phase of the emitted radiation. More generally one studies the asymptotic behavior of integrals of the type $\int_M a(y)e^{ikf(y)}dy$, where M is a Riemannian manifold with volume element dy and

---

[1] This is what the algebraists would call a functorial approach to the theory, very much in the sense of Grothendieck's algebraic geometry.

$a(y), f(y)$ are suitable functions on M. If $C_f := \{s \in M : (df)_s = 0\}$ is the critical set, i.e., the points in M where the gradient of f vanishes, and if $H(s) = (d^2 f)_s$ denotes the matrix associated to the Hessian of f, i.e., the matrix of the quadratic form defined by the second order terms in the Taylor expansion of f about the point s, then under suitable assumptions on the data, e.g., f is a Morse function, so that C_f is a finite set of isolated points, the principle of stationary phase can be stated by the beautiful formula

$$\int_M a(y)e^{2\pi i k f(y)}dy = \{\sum_{s\in C_f} e^{2\pi i sgn H(s)/8}\frac{e^{2\pi i k f(s)}}{\sqrt{|det\ H(s)|}}\}k^{-n} + O(k^{-n-1}).$$

For a precise statement and proof see Hormander ([6], p.220) or Sternberg and Gulliman ([5], p.6). The central theme of this paper is that the method of stationary phase for finite rings is at the heart of the explicit evaluation of exponential sums.

Contents: In §2 we review the elementary properties of quadratic Gauss sums. In §3 we recall some results of Kummer on Gauss sums over residue rings which do not seem to be as well known as his work related to Gauss sums over finite fields. In §4 we introduce the structure theory of finite rings and describe in some special cases the theory developped systematically by Lamprecht and Dwork. In §5 we introduce some elementary notions from number theory which seem to have been first discovered by L.K. Hua and which play an important role in additive number theory (the circle method). This method of Hua can be viewed as a primitive form of Igusa's method of p-adic stationary phase which is explained in §6.

2 Quadratic Gauss Sums

By any measure of importance the quintessential Gauss sum is the quadratic one

$$G(n) := \sum_{s\in \mathbf{Z}/n\mathbf{Z}} e^{\frac{2\pi i s^2}{n}} = \frac{1+(-i)^n}{1-i}\cdot\sqrt{n}, \tag{1}$$

whose original evaluation by Gauss was a long and arduous task lasting over several years. In the sum defining G(n), s runs over any complete system of residues modulo n. One of the consequences of the above evaluation is the proof of the Quadratic Reciprocity Law. In fact, after some elementary manipulations the two statements are equivalent. The atraction which these two results excercised during the nineteenth century is best measured by the number of purported proofs published by 1899- over 150 ! In his *Elementary Number Theory Book*, Landau gave four distinct proofs of the evaluation of G(n) with great clarity and insightfullness which are still a pleasure to read. The first proof is essentially due to Dirichlet and in modern notation is based on the distributional identity (Poisson Summation Formula): $\delta_0 = \sum_{s\in \mathbf{Z}} e^{2\pi i s}$, applied to the function $f_s(\xi) = e^{2\pi i\frac{(s+\xi)^2}{n}}$, where δ_0 is Dirac's measure at 0. Its great merit stems from the fact that is it is insensitive to the parity of n. The second proof, that of Kronecker, is based on the Cauchy residue theorem and, like the first, has an adelic generalization. The third proof, that of I. Schur, is based on the spectral properties of the matrix $A = (e^{\frac{2\pi i s r}{n}})$ and is of course well known to those familiar with the discrete Fourier transform. The last proof is due to Mertens and is based on some elementary manipulations using the Chinese Remainder theorem and the basic properties of the finite ring $R = \mathbf{Z}/n\mathbf{Z}$. As Merten's proof exhibits the seeds of the essential structure of the general theory of finite rings needed to study Gauss sums, we recall some of the main steps. In the definition of the Gauss sum G(n)in (1) we replace the quadratic expression s^2 by ms^2 and denote the new sum by $G(m,n)$. Note that $G(m,1) = 1$. For $n > 1$ and $n = \prod_{p|n} p^\ell$ its canonical decomposition, an iteration of the Chinese Remainder Theorem yields the multiplicative relation $G(1,n) = \prod_{p|n} G(\frac{n}{p^\ell}, p^\ell)$. The next crucial step is the identity $G(m,p^\ell) = pG(m,p^{\ell-2})$, which is valid for $\ell > 2$ and $p \not| 2m$. Its derivation uses the parametrization of the residue classes modulo n by the expression $p^{\ell-1}z+t$ in which z is allowed to run over residue classes modulo $p^{\ell-1}$ and t over any complete set of residue classes modulo p. Iteration of the last identity yields the result $G(m,p^\ell) = p^{\frac{\ell}{2}}$ for ℓ even, and $G(m,p^\ell) = p^{\frac{\ell-1}{2}}G(m,p)$ for ℓ odd. Putting together these properties we see that the evaluation of the Gauss sum $G(1,n)$ is then reduced to the evaluation of the sums $G(m,p)$ for odd primes p which do not divide 2m. For the latter one proves in an elementary way that $G(m,p)^2 = (-1)^{\frac{p-1}{2}}p$, and the evaluation is reduced to the determination of a sign. As Landau states, "*this is where the computation really begins in earnest* " (See Landau [2], p.216). What Mertens proof brings out very clearly is the need for some sort of "primary decomposition " for a finite ring R which will replace the use of the Chinese Remainder Theorem, and secondly some sort of filtration on the set of ideals which mimics the parametrization "$p^{\ell-1}z + t$". As will be seen later on, it is the systematic use of Jordan-Holder composition series on Finite rings with some sort of Gorenstein property which leads to similar results for general Gauss sums over fairly arbitrary finite rings. This is in essence the theory of Lamprecht and Dwork and will be explained in §4. In the next section we describe the general results obtained by Kummer for a finite primary ring of the type $R = \mathbf{Z}/p^\ell\mathbf{Z}$.

3 Kummer's work on general Gauss sums

The results described in this section are due to Kummer and appeared in his paper [13]; to the best of my knowledge they were never used in any serious way in his work on cyclotomic extensions and Fermat's Last Theorem and they seemed to have attracted little notice. For our purposes their relevance stems from the appearance of a p-adic approximation technique which is at the heart of the method of stationary phase.

Throughout this section we deal exclusively with the residue ring $R = \mathbf{Z}/p^\ell\mathbf{Z}$

with p and odd prime. We recall that the multiplicative group $R^{\times}$ is cyclic of order $\varphi(p^{\ell})$ and we fix a generator (primitive root) which we denote by g. We fix a generic additive character $\psi : R^{+} \to \mathrm{C}^{\times}$ so that the Pontriyagin dual of R^{+}, viewed as an R-module has the Gorenstein property, i.e., $Hom(R^{+}, \mathrm{C}^{\times}) \simeq R \cdot \psi$. Let $\chi : R^{\times} \to \mathrm{C}^{\times}$ be a multiplicative character. Let $r \in R$ be an element that satisfies the relation $\chi(g)^{p-1} = \psi(r(p-1))^{p}$. Let e' be an integer satisfying the relation $g^{p-1} = 1 + e'p$. If r is not a multiple of p, we define an integer ρ by solving the exponential equation

$$r(p-1) + \frac{g^{\rho}}{p}\Big(e'p - \frac{e'^2p^2}{2} + \frac{e'^3p^3}{3} - \cdots - \frac{(-1)^{\ell'-1}e'^{\ell'}p^{\ell'}}{\ell'}\Big) \equiv 0 \text{ mod. } p^{\ell'},$$

where $\ell' = [\frac{\ell}{2}]$. We recognize here the presence of the p-adic logarithm:

$$r(p-1) + \frac{g^{\rho}}{p} log_p(1 + e'p) \equiv 0 \text{ mod. } p^{\ell'}.$$

See (Iwasawa [14], p. 38). The Gauss sum is defined by

$$G(\chi, \psi) := \sum_{s \in R^{\times}} \chi(s)\psi(s).$$

The following are the main results of Kummer disguised in a different notation.

THEOREM. *With notations and assumptions as above, we have*

- $G(\chi, \psi) = 0$ *if* $r \equiv 0$ *mod.* p,

Suppose $r \not\equiv 0$ *mod.* p*. Then*

- $|G(\chi, \psi)|^2 = p^{\ell}$, *and*
- $G(\chi, \psi) = \begin{cases} p^{\ell/2}\chi(g^{\rho})\psi(g^{\rho}) & \text{if } \ell \text{ is even} \\ G(\psi)p^{\frac{\ell-1}{2}}\chi(g^{\rho})\psi(g^{\rho}) & \text{if } \ell \text{ is odd,} \end{cases}$

where

$$G(\psi) = \sum_{s \in \mathbf{Z}/p\mathbf{Z}} \psi'(s^2), \qquad \psi'(s) = \psi(\frac{r}{2}p^{\ell-1}s).$$

We end this review of Kummer's contributions by also mentioning the very important special cases of Stickelberger's theorem discovered by Kummer.

4 Gauss Sums over Finite Rings

Let R be a finite commutative ring with identity. If S is a subset of R, we denote by $N(S)$ its cardinality. Let $R^{\times}$ denote the multiplicative group of invertible elements in R. Given an additive character $\psi : R^+ \to \mathbf{C}^\times$ and a multiplicative character $\chi : R^\times \to \mathbf{C}^\times$, we define as before the Gauss sum $G(\chi, \psi) := \sum_{s \in R^\times} \chi(s)\psi(s)$, and note that its value is always an algebraic integer. Because of the finiteness condition, the structure theory of such rings has been known for a long time (See Krull [11]). We recall here some of their most salient properties which are of relevance to the study of $G(\chi, \psi)$. Recall that a ring R is called *primary* when its zero divisors are all nilpotent; in this case there is a unique prime ideal. The Noether-Lasker theorem gives a direct product decomposition

$$R = R_1 \oplus \cdots \oplus R_n,$$

where each component R_i is primary. The identity element in R has a decomposition $1_R = 1_1 \oplus \cdots \oplus 1_n$ with the $1_i \in R_i$ satisfying $1_i^2 = 1_i$, $1_i \cdot 1_j = 0\,(i \neq j)$. The Pontryagin duals $Hom(R^+, \mathbf{C}^\times)$ and $Hom(R^\times, \mathbf{C}^\times)$ are also endowed with similar decompositions. More precisely, if an element $s \in R$ has a representation $s = s_1 + \cdots + s_n$, with $s_i \in R_i$, we write for $\psi \in Hom(R^+, \mathbf{C}^\times)$, $\psi(s) = \prod_{i=1}^n \psi_i(s_i)$, where $\psi_i(s_i) = \psi(1_i s)$ is an additive character of R_i. Similarly for $\chi \in Hom(R^\times, \mathbf{C}^\times)$ we write $\chi(s) = \prod_{i=1}^n \chi_i(s_i)$, where $\chi_i(s_i) = \chi(s_i^*)$ and $s_i^* = s_i + \sum_{k \neq i} 1_k$. Note that the group $\tilde{R}_i^\times$ generated by the s_i^* is isomorphic to $R_i^\times$. With these notations it is not difficult to show that the Noether-Lasker decomposition induces a product factorization for the Gauss sum:

$$G(\chi, \psi) = \prod_{i=1}^n G_i(\chi_i, \psi_i),$$

where each of the factors $G_i(\chi_i, \psi_i)$ is a Gauss sum associated to R_i. This then reduces the study of Gauss sums to those defined on primary rings.

Suppose now that R is a finite primary commutative ring and let $\wp$ denote its unique prime ideal with residue class field $\mathbf{F}_\wp = R/\wp$ of cardinality $q = p^f$. We now introduce a condition on R which is fundamental in the study of Gauss sums:

Condition (G)[2]: The Pontryagin dual of R^+ is a rank one R-module.

This condition guarantees that we can find an additive character ψ such that $Hom(R^+, \mathbf{C}^\times) \simeq R \cdot \psi$, i.e., any other character ψ' of R^+ has the form $\psi'(s) = \psi(r \cdot s)$ for some $r \in R$. Such a character ψ will be called generic. If $\underline{q} \nsubseteq R$ is a primary ideal in R, and $\underline{q} \subseteq \wp$, then all $s \in R$ satisfying $s \equiv 1$ mod. $\underline{q}$ are nonzero divisors and hence $s \equiv$ mod. $\wp$ and $s \in R^\times$. We say that a multiplicative character χ is defined modulo $\underline{q}$ if $\chi(s) = 1$ for all $s \equiv$ mod. $\underline{q}$. The conductor $\mathcal{F}_\chi$ of a character χ is defined as the largest common divisor of all ideals $\underline{q}$ with respect to which χ is defined modulo $\underline{q}$. When the conductor $\mathcal{F}_\chi = (0)$, we say that χ is proper. The first main result of the theory is that unless R satisfies the condition (G), there do not exist proper multiplicative characters χ. A simple substitution in the definition of $G(\chi, \psi)$ shows that if R does not satisfy condition (**G**) and $\psi : R^+ \rightarrow \mathbf{C}^\times$ is an additive character with kernel $\underline{q}_\psi := \{s \in R : \psi(sx) = 1 \text{ for all } x \in R\}$, then the existence of a single element $r \equiv 1$ mod. $\underline{q}_\psi$ with $\chi(r) \neq 1$ implies $G(\chi, \psi) = 0$. Furthermore, if $\chi(r) = 1$ for all $r \equiv$ mod. $\underline{q}_\psi$, then one shows that

---

[2] G stands for Gorenstein.

$$G(\chi, \psi) = N(\underline{q}_\psi) \sum_{s \in R^\times, s \text{ mod.} \underline{q}_\psi} \chi(s)\psi(s).$$

Thus the essential problem becomes that of calculating Gauss sums associated to primary rings that satisfy condition (**G**).

DEFINITION. *A finite primary ring R is called reducible if its zero ideal can be expressed as the intersection of proper primary ideals; otherwise it is called irreducible.*

It is not difficult to show that the irreducible rings are precisely the ones that satisfy the condition (G). If $\underline{q}$ is an ideal in R, we denote its annihilator by $ann_R(\underline{q}) := \{s \in R : s\underline{q} = (0)\}$. In a primary ring R, the operation $\underline{q} \mapsto ann_R(\underline{q})$ defines an important permutation of the family of ideals. If $\wp$ is the prime ideal of an irreducible primary ring R, then the integer $n = n(R)$ such that $\wp^n = (0)$ but $\wp^{n-1} \neq (0)$ is called the exponent of the ring R. We note that $ann_R(\wp) = \wp^{n-1}$ and $ann_R(\wp^{n-1}) = \wp$. The length $m = m(R)$ of Jordan-Holder composition series (c.s.) for a finite irreducible ring R

$$(0) = \underline{q}_0 \subset \underline{q}_1 \subset \underline{q}_2 \subset \cdots \subset \underline{q}_{m-1} = \wp \subset \underline{q}_m = R,$$

with

$$\underline{q}_{i+1}/\underline{q}_i \simeq \mathbf{F}_\wp, \qquad 0 \leq i \leq m-1$$

is called the length of the composition series. Note that Card(R)=$q^{m(R)}$.

DEFINITION. *If the two ideals $\underline{r} := ann_R(\underline{q}) \subset \underline{q}$ appearing in a (c.s.) for a finite irreducible ring R have the property that no element $s \in R$ satisfies $s \equiv 0$ mod.$\underline{q}$, $s \not\equiv 0$ mod. $\underline{r}$ and $s^2 = 0$, then they are called a pair of characteristic ideals of the ring.*

The following structure theorem was proved by Lamprecht ([10], p.160) and to some extent by Dwork (see Thesis, Chap. III). It is a key ingredient in the evaluation of Gauss sums.

THEOREM. *Let R be a finite irreducible primary ring with prime ideal $\wp$ and residue field $R/\wp = \mathbf{F}_\wp$ with q elements. Then R has a pair of characteristic ideals $\underline{q}$ and $ann_R(\underline{q})$; such a pair satisfies $\wp \cdot \underline{q} \subseteq \underline{r}$, $\underline{r}^2 = (0)$ and the structure of the ring satisfies one of the following cases:*

- *If $m(R) = 2\mu + 1$, then $(\underline{q}) = q^\mu$, $\underline{q}/\underline{r} \simeq \mathbf{F}_\wp$ as modules and a given representation $\underline{q} = \underline{r} + (\omega)$ satisfies $\underline{q}^2 = \wp^{n-1} = (\omega^2) \neq (0)$.*

- *If* $m(R) = 2\mu$, *then* $N(\underline{q}) = q^{\mu}$ *and hence* $\underline{q} = \underline{r}$, $\underline{q}^2 = (0)$. *(This occurs only when* $char(\mathbf{F}_{\wp}) = 2$*).*
- *If* $m(R) = 2\mu$ *and* $char(\mathbf{F}_{\wp}) \neq 2$, *then* $N(\underline{q}) = q^{\mu-1}$, $\underline{q}^2 = \wp^{n-1}$, $\underline{q}/\underline{r} \simeq \mathbf{F}_{\wp} \oplus \mathbf{F}_{\wp}$ *as modules and a given representation* $\underline{q} = \underline{r} + (\omega_1) + (\omega')$ *satisfies* $\omega\omega' = 0$. $\omega^2 \neq 0$, $\omega'^2 \neq 0$; *there is a relation* $\omega^2 + \alpha\omega'^2 = 0$ *with* α *not a square modulo* $\wp$.

DEFINITION. *A Jordan-Holder composition series* $(0) = \underline{q}_0 \subset \underline{q}_1 \subset \cdots \subset \underline{q}_m = R$ *in a finite irreducible ring* R *with the property that each ideal* $\underline{q}_i$ *and its corresponding annihilator* $ann_R(\underline{q}_i) = \underline{q}_{m-i}$ *is said to be inversion invariant (i.i.c.s.).*

It can be shown that for many rings R of interest in applications it is always possible to find (i.i.c.s.). For instance if the residue field $\mathbf{F}_{\wp}$ is of characteristic 2, an (i.i.c.s.) always exists. We are now in a position to state one of the main results of the theory which asserts that under natural hypothesis on the ring R, in so far as the evaluation of Gauss sums is concerned, the analogues of Kummer's results hold. More precisely, suppose R is a finite primary ring. Let $\underline{q}$ and $\underline{r} = ann_R(\underline{q})$ be a characteristic pair with $\underline{r}^2 = (0)$. Suppose $\chi(1+s) = \psi(-\varepsilon_\chi s)$ for all $s \in \underline{r}$ for some $\varepsilon_\chi \in R^\times$ (uniquely determined modulo $\underline{q}$).

THEOREM. *With notations as above we have*

- *If* χ *is not proper,* $G(\chi, \psi) = 0$.

Suppose χ *is proper and* $\underline{q}$, $\underline{r} = ann_R(\underline{q})$ *is a characteristic pair in an (i.i.c.s.). Then*

- $|G(\chi, \psi)|^2 = N(R)$.
- $G(\chi,\psi) = \begin{cases} N(\underline{q})\chi(\varepsilon_\chi)\psi(\varepsilon_\chi) & \textit{if } m(R) \textit{ is even} \\ G_{\underline{q}}(\chi,\psi)N(\underline{q})\chi(\varepsilon_\chi)\psi(\varepsilon_\chi) & \textit{if } m(R) \textit{ is odd,} \end{cases}$

where

$$G_{\underline{q}} = \sum_{s \in \underline{q}/\underline{r}} \chi(1-s)\psi(-\varepsilon_\chi s).$$

REMARK. Lamprecht and Dwork make a very precise analysis of the sums $G_{\underline{q}}(\chi, \psi)$ and show that it is related to the theta-type sum

$$G(\Delta) = \sum_{s \in \underline{q}/\underline{r}} \Delta(s),$$

constructed with a function $\Delta : R \to \mathbf{C}^\times$ that satisfies the relation

$$\Delta(s)\Delta(s') = \Delta(s+s')\psi(-\varepsilon_\chi w^2 ss')$$

for some suitable element $w \neq 0$. This is the most delicate part of the work and in a fuller version of this article we will give an alternative derivation of the results.

The results of Lamprecht and Dwork have important applications to the root numbers that appear in the functional equation of Artin L-functions. A deep and significant result of Langlands and Dwork is that such root numbers have a representation as a product of local root numbers which in the abelian case coincide with Gauss sums over finite rings. The original proof of this result used in a fundamental way the four relations satisfied by Gauss sums over finite fields: Hasse-Davenport relation, Hasse's distribution relation, Dwork's formula, and Langlands relation. In the case of Gauss sums associated to finite rings one also has a Hasse-Davenport relation which generalizes the one already proved by Dwork. We believe that analogues of the other three identities hold, but this needs some further work. Other results on Gauss sums over finite rings will be included in the final version of this paper.

5 Arithmetic Stationary Phase

In his work on the additive theory of prime numbers, L.K. Hua was lead to a refinement of several tools that had made their appearance in earlier versions of the circle method. One of these, the Fundamental Lemma, deals with exponential sums and merits special mention in this report because it is in all its essentials a primitive one-variable version of the method of stationary phase which will be explained in the next section.

In describing Hua's results, we shall use the following notation: $R_\ell = \mathbf{Z}/p^\ell\mathbf{Z}$, for additive character we shall use $\psi_\ell(s) = e^{2\pi i s/p^\ell}$, $f(x) = b_0 + b_1x + b_2x^2 + \cdots + b_kx^k$ will be a polynomial of degree k with integer coefficients; note that without loss of generality we may assume that the constant coefficient $b_0 = 0$. Hua's method deals with exponential sums of the type

$$G(p^\ell, f) := \sum_{s\in R_\ell} e^{2\pi i f(s)/p^\ell}.$$

The study of the more general sum in which P^ℓ is replaced by an arbitrary integer q is reduced by the Chinese Remainder Theorem to the above type. The special case $\ell = 1$, in which case the ring R_1 is just the finite field of p elements is dealt with adequately by the Carlitz-Uchiyama bound(Weil's Riemann Hypothesis). In the following we shall assume that $\ell > 1$. We introduce first a definition.

DEFINITION. *A polynomial congruence* $f(x) = b_0 + b_1x + \cdots + b_kx^k \equiv 0\,(mod.\,p)$ *has a root of multiplicity* m *at* μ *when* $f(x) = (x-\mu)^m f_1(x) + pf_2(x)$ *with* f_1, f_2 *polynomials such that* $p \not| f_1(\mu)$.

We assume that $p \not| (b_1, b_2, \ldots, b_k)$. Let a be the exact power of p which devides the coefficients of the derivative of f, namely $p^a \| (b_1, 2b_2, \ldots, kb_k)$. We define the critical set for the polynomial f by

$$C_f(p) := \{s : p^{-a} f'(s) \equiv 0 \text{ mod. } p\}.$$

Observe that if the multiplicity of an element $\mu \in C_f(p)$ is denoted by m_μ, then the sum of the $m'_\mu s$ is at most k-1 by Lagranges theorem. Given an element μ in the critical set $C_f(p)$ we consider the polynomial $f(px+\mu) - f(\mu) = c_1 x + \cdots + c_k x^k$. Its coefficients are related to those of f by the triangular linear system $c_i = p^i \sum_{h=i}^k b_h \begin{pmatrix} h \\ i \end{pmatrix} \mu^{h-i}$, $(i = 1, \ldots, k)$. We let $\rho(\mu)$ be the largest power of p which devides all the coefficients c_i, that is $p^{\rho(\mu)} \| (c_1, \ldots, c_k)$. Clearly $\rho(\mu) \geq 1$. We associated to μ and $\rho(\mu)$ a new polynomial of degree k:

$$f_\mu(x) := p^{-\rho(\mu)}(f(px+\mu) - f(\mu)).$$

For a critical point μ, we denote by $p^{\sigma(\mu)}$ the highest power of p which devides the coefficients of the derivative $f'_\mu(x)$; it is clear that $1 \leq \sigma(\mu) \leq k$. The main ingredients in Hua's proof of the Fundamental Lemma can now be stated in the following form.

LEMMA. *With notations and assumptions as above, we have*

$$\sum_{s \in R_\ell} \psi_\ell(f(s)) = \sum_{\mu \in C_f(p)} \psi_\ell(f(\mu)) p^{\rho(\mu)-1} \sum_{r \in R_{\ell-\rho(\mu)}} \psi_{\ell-\rho(\mu)}(f_\mu(r)).$$

Furthermore, if a critical point $\mu \in C_f(p)$ has multiplicity m_μ, then the number of solutions of the congruence $p^{-\sigma(\mu)} f'_\mu(x) \equiv 0 \,(mod.\, p)$, counting multiple solutions with their multiplicities, is at most m_μ; that is, the critical set for the polynomial $f_\mu(x)$ satisfies $Card\, C_{f_\mu}(p) \leq m_\mu$.

The proof of this lemma is given in Hua ([3], §4, Lemma 1.4 and §5, Lemma 1.6). For a very clear and lucid presentation of the argument see Vaughan ([4], §7.2). The central ideas that lie behind Hua's results and which have further applicability to exponential sums over rings which are more general than R_ℓ are again of three types:(i) some sort of factorization of the sum which reflects the primary decomposition of the ring; in Hua's case this is accomplished simply by the Chinese Remainder Theorem. (ii) some sort of iteration whereby a sum over the ring R_ℓ is related to sums over smaller rings $R_{\ell'}$ with $\ell' < \ell$; this is possible in general whenever there is a Jordan Holder decomposition series as is brought out very clearly in Lamprecht's work on Gauss sums. (iii) some sort of "*Arithmetic Morse Lemma*" in which iteration of the map $f \mapsto f_\mu$ corresponds to a change of coordinates near the critical point μ so that the end result is a quadratic function in the new variables. (For the classical Morse Lemma see Guillemin and Sternberg [5], Appendix I, p. 16). To exemplify these ideas we mention a particular case which is easy to prove directly: Suppose the polynomial $f(x)$ is of Morse type which simply means that its critical set $C_f(p)$ consists of simple points, and that its degree $deg(f) < p$, then we have

$$\sum_{s \in \mathbf{Z}/p^\ell \mathbf{Z}} e^{2\pi i f(s)/p^\ell} = p \sum_{\mu \in C_f(p)} e^{2\pi i f(\mu)/p^\ell} \Big\{ \sum_{r \in \mathbf{Z}/p^{\ell-2}\mathbf{Z}} e^{2\pi i \{\frac{1}{2} f''(\mu) r^2 + \cdots\}/p^{\ell-2}} \Big\}.$$

Here we see easily that in the even case ($\ell = 2\ell'$), after ℓ' iterations, the inner sum disappears and the value of the sum takes a very simple form. In the odd case ($\ell = 2\ell' + 1$), the iteration stops after ℓ' steps and the inner sum is now our old friend, the quadratic Gauss sum $G(p, g(x))$ constructed with the quadratic function $g(x) = \frac{1}{2}f''(\mu)x^2$. In the most general situation it is clear that one always has a representation of the form

$$\sum_{s\in R_\ell} e^{2\pi i f(s)/p^\ell} = \sum_j a_j(\sqrt{p})^{w(j)},$$

where the complex numbers a_j have absolute value 1 and the exponents $w(j)$ are integers which in the "Morse Case" lie in the set $\{0, 1, \dots, \ell\}$. This applies particularly to the polynomial $f(x) = x^k + ax$ which plays such an important role in the treatment of the major arcs in Waring's problem and in the proof of Vinogradov's Mean Value Theorem (see Vaughan [4], Chap. 4).

6 P-adic Stationary Phase

The method of stationary phase has a p-adic analogue which has been studied by Igusa in connection with his investigation of the validity of a generalized Poisson Summation Formula. It has a strong structural resenblance with the classical case and with Laumon's stationary phase for ℓ-adic sheaves in algebraic geometry.

Let K be a non-archimedean field of characteristic zero, e.g., a finite algebraic extension of $\mathbf{Q}_p$, with exponential valuation $ord : K \longrightarrow \mathbf{Z}$ and residue class field $\mathbf{F}_q$. Let $|x|_K = q^{-ord(x)}$ be the corresponding absolute value. Let P be the maximal ideal in the ring of integers R of K, and let π be a local uniformizing parameter so that $P = \pi R$ and $\mathbf{F}_q = R/P$. Suppose V is a smooth R-scheme [3] of relative dimension n and f a function on V (i.e. f is an R-morphism from V to the affine line $\mathbf{A}^1$). We suppose that f is a Morse function in the sense that thecritical set

$$C_f := \{s \in V : grad_s f = 0\}$$

is finite and finite and etale over R. Let $\psi \in Hom_c(K^+, \mathbf{C}^\times)$ be a non-trivial additive character of K. It is customary to take ψ to be the canonical character defined by $\lambda_p : \mathbf{Q}_p \longrightarrow \mathbf{Q}_p/\mathbf{Z}_p \longrightarrow \mathbf{Q}/\mathbf{Z} \xrightarrow{e^{2\pi i(\cdot)}} \mathbf{T}$ composed with the trace map $Tr_{K/\mathbf{Q}_p}$. Let $Phi \in \mathcal{S}(V(K))$ be a Scwartz function on the K-points of V. We define a function F^*_Φ on K by

$$F^*_\Phi(i^*) = \int_{V(K)} \Phi(x)\psi(i^* f(x))dx, \qquad (i^* \in K).$$

---

[3] In accordance with the low brow character of our exposition, we may restrict our considerations to an algebraic closed set V defined by a finite number of polynomial equations with coefficients in R which is free of singularities, e.g., affine n-space $\mathbf{A}^n$, or projective n-space $\mathbf{P}^n$.

This is the analogue of a Fourier integral. To obtain information on F^*_Φ one needs to make certain assumptions regarding the algebraic set V and the function f. An example of one of the fundamental results in the theory concerns the case when $V = \mathbf{A}^n$ is affine n-space and Φ is the characteristic function of the set $\mathbf{A}^n(R)$: if f is a polynomial in $K[x_1, \ldots, x_n]$, then

$$(1) \qquad F^*_\Phi(i^*) = q^{-nm} \sum_{i \in \mathbf{A}^n(R/P^n)} \psi(i^* f(i)), \qquad i^* = u \cdot \pi^{-m},$$

where $u \in R^\times$ and $q = chard(R/P)$.

THEOREM. (Igusa[7]) *With notations and assumptions as above, there exist a finite set of complex numbers* $\{\lambda_j\}$, *and a positive integer* e_0 *such that for* $m \geq e_0$,

$$(2) \qquad F^*_\Phi(i^*) = \sum_{j=1}^{k} a_j \chi_j(u) |i^*|_K^{-\lambda_j} (log|i^*|_K)^{m(j)}, \qquad i^* = u \cdot \pi^{-m},$$

where the a_j *are complex constants,* $\chi_j \in Hom_c(R^\times, \mathbf{C}^\times)$ *are characters of conductor a power of* P, $Re(\lambda_j)$ *are positive real numbers, and the* $m(j)$ *are integers* $< n$.

REMARKS. 1. The existence of such asymptotic expansions is equivalent to the stable behaviour of the singular series: for $i \in R$, the quotient $N_m(i)/q^{(n-1)m}$ is independent of m for m sufficiently large, where

$$N_m(i) := Card\{x \in \mathbf{A}^n(R/P) : f(x) \equiv i \text{ mod. } P^m\}.$$

2. The original proof of Igusa requires Hironaka's resolution of singularities, hence the restriction to characteristic zero; the parameters a_j, λ_j, χ_j and $m(j)$, and especially e_0 can be made explicit by using the numerical data which enters into Hironaka's resolution of embedded singularities.

3. From the point of view of applications, especially to error control codes over the finite ring R/P^ℓ, one can consider the combination of (1) and (2) as a generalization of the Carlitz-Uchiyama bound.

There are special cases where one can be more precise. For instance, when the critical set $C_f(R)$ is empty, $F^*_\Phi(i^*)$ vanishes for $m \geq 2$ and Φ the characteristic function of the set $V(R)$. When the critical set $C_f(R)$ is not empty and $i^* = u\pi^{-m}$ with $m \geq 2$, the Fourier integral is a sum of local terms indexed by the finite number of elements in $C_f(R)$. In fact, for m even, we have

$$\int_{V(R)} \psi(i^* f(x)) dx = \Big\{ \sum_{s \in C_f(R)} \psi(i^* f(s)) \Big\} |i^*|_K^{-n/2}.$$

For the case $i^* = u\pi^{-m}$, $m \geq 3$ odd, we have

$$\int_{V(R)} \psi(i^* f(x)) dx = \Big\{ \sum_{s \in C_f(R)} \zeta_8^{sgn\, H(s)} \psi(i^* f(s)) \Big\} |i^*|_K^{-n/2},$$

where $H(s)$ is the Hessian matrix $(d^2f)_s$ of f at the critical point $s \in C_f(R)$, and $\zeta_8^{sgn\, H(s)}$ is just a symbolic way of writing the fourth root of unity whose value is given by the multidimensional Gauss sum

$$\zeta_8^{sgn\, H(s)} = q^{-n/2} \sum_{z \in (R/P)^n} \psi(\pi^{m-1} i^* H(z)).$$

The seemingly simpler sum

$$\sum_{z \in V(R/P)} \psi(i^* f(z)), \qquad i^* = u\pi^{-1},\, u \in R^\times,$$

is by far the most difficult case and information is available only under very special assumptions on f, e.g., f is a polynomial whose highest homogeneous term is non-singular (by Deligne). The purity and decomposition theorems in intersection cohomology and the theory of perverse sheaves yields useful information in more general situations.

The work of Igusa has been carried quite far by Denef and his collaborators. His most notable results concern explicit formulas for the Igusa zeta functions which in turn translate into very precise formulas for exponential sums (in one and several variables) over residue rings arising as the quotient of a local ring of integers by powers of its maximal ideal. Lack of space prevents us from describing this very interesting aspect of the theory. The interested reader can consult Denef's Bourbaki Report ([8]).

Summary. In conclussion we like to mention that for all practical purposes, a great deal of knowledge about exponential sums is now known, and any problem involving such sums is either reducible to one of the cases reviewed in this survey or to an exponential sum over a finite field in which case the methods of Deligne, Katz, Gaber,Adolphson, Sperber,..., are applicable. What is important now is to find applications of these results to practical problem in the same way that the Carlitz-Uchiyama bound and similar estimates found in the construction of error control codes, digital signal processing and discrete Fourier transforms. We now have a very satisfactory construction of algebraic geometric codes over rings due to Judy Walker. As a first approximation, the challenge that lies in front of us is to find suitable generalizations of the work MaCliece concerning the weight distribution of code words in cyclic codes in the new setting of algebraic geometric codes over finite rings !

Bibliography

1. **R. E. Blahut**, *Algebraic Fields, Signal Processing, and Error Control*, Proc. of IEEE, vol. 73, no. 5 (1985), pp. 872-893.

2.**E. Landau**, *Elementary Number Theory*, Chelsea Publishing Co., New York, 1966.

3.**L. K. Hua**, *Additive Theory of Prime Numbers*, American Mathematical Society, Providence, R.Island, 1965.

4.**R. C. Vaughan**, *The Hardy-Littlewood Method*, Cambridge University Press, Cambridge, England, 1981.

5.**V. Guillemin and S. Sternberg**, *Geometric Asymptotics*, American Mathematical Society, Providence, R. Island, 1977.

6.**L. Hormander**, *The Analysis of Linear Partial Differential Operators I*, Springer-Verlag, New York, 1983.

7. **J.-I. Igusa**, *Lectures on Forms of Higher Degree*, Tata Institute of Fundamental Research, Bombay, 1978.

8. **J. Denef**, *Report on Igusa's Local Zeta Function*, Seminaire Bourbaki, n^o 741, Asterisque 201-203(1991), pp. 359-386.

9. **B. Dwork**, *On the Root Number in the Functional Equation of the Artin-Weil L-Function*, (Thesis, Columbia University, July 19, 1954).

10.**E. Lamprecht**, *Allgemeine Theorie der Gaussschien Summen in endlichen kommutativen Ringen*, Math. Nachr., vol. 9 (1953), pp. 149-196.

11.**W. Krull**, *Algebraische Theorie der Ringe. I;II;III*, Math. Ann. 88(1923),pp. 80-122; 91 (1924), pp. 1-46; 92(1924), pp. 183-213.

12.**J. L. Walker**, *Algebraic Geometric Codes Over Rings*, (Thesis, University of Illinois, 1995.

13. **E. Kummer**, *Note sur une expression analogue a la resolvante de Lagrange pour la equation* $z^p = 1$, Atti dell'Accademia Pontifica de Nuovi Lincei VI, (1852-1853), pp. 237-241.

14.**K. Iwasawa**, *Lectures on p-Adic L-Functions*, Annals of Math. Studies, Princeton Univ. Press, vol. 74, Princeton, N.J., 1972.

Exponential Sum Estimates and Applications

Wen-Ching Winnie Li*
Department of Mathematics, Pennsylvania State University
University Park, PA 16802

Abstract

In this paper we review certain character sum estimates over Galois rings obtained in [8] with the underlying curve being the projective line. As applications, we show how such estimates are used to construct periodic sequences of low correlation as well as codes with good parameters.

§1 Introduction.

Character sum estimates have been used substantially in coding theory, in particular, for weight enumeration and for constructing families of periodic sequences with low cor-relation. Typically one considers character sums of the following form:

$$\sum_x \psi(f(x)), \quad \sum_x \chi(g(x)), \quad \text{and} \quad \sum_x \psi(f(x))\chi(g(x)).$$

A. Weil [16] was the first to give estimates of the first two kinds where f and g are rational functions on the projective line $P^1(k)$ over a finite field k with q elements, ψ is a nontrivial additive character of k, χ is a nontrivial multiplicative character of k, and the sum is over points $x \in P^1(k)$ at which f and/or g have no poles and the vaules $f(x)$, $g(x)$ fall in the domain of the character(s) involved. An estimate of the first sum for curves over k of arbitrary genus was obtained by Bombieri [1]. Since the paper [4] by Hammons et al., coding theorists got interested in such character sums over points defined over $\mathbb{Z}_4$, or more generally, over Galois rings, which we explain presently. Denote by p the characteristic of k. Let K be the unramified extension of the p-adic field $\mathbb{Q}_p$ with k being its residue field. The ring of integers $\mathcal{O}$ of K divided by $p^r\mathcal{O}$ for $r \geq 1$ is a Galois ring and an! y Galois ring arises this way. Let C be the variety defined by the equation $x^q = xz^{q-1}$. Thus $P^1(k) = C(k)$ and it will be replaced by the lifting to points over K, namely, $C(K)$. The character ψ is a character of $\mathbb{Z}/p^r\mathbb{Z}$ of order p^r composed with the trace map from $\mathcal{O}/p^r\mathcal{O}$ to $\mathbb{Z}/p^r\mathbb{Z}$. Coding theorists obtained various estimates for sums of the first kind and third kind with f a polynomial in x and $1/x$ over $\mathcal{O}/p^r\mathcal{O}$, and g a simple polynomial over $\mathcal{O}/p\mathcal{O}$. See the review paper by Helleseth et al. [6] and the references therein. In a recent paper [8], we derived estimates of the above three kinds under the most general situation, namely, f and g rational functions over $\mathcal{O}$, ψ an additive character of K nontrivial on $\mathcal{O}$ and χ a multiplicative character of K nontrivial on the group of units in $\mathcal{O}$. Of course, when ψ is trivial on $p^r\mathcal{O}$ (re! sp. χ trivial on $1 + p^r\mathcal{O}$), then f (resp. g) may be regarded as a rational function over the Galois ring $\mathcal{O}/p^r\mathcal{O}$. We shall review main results in §2.

---

*Research supported in part by a grant from the National Science Foundation no. DMS-9622938

The aforementioned character sums are majorized by $M\sqrt{q}$, as a consequence of the Riemann hypothesis for curves over k proved by Hasse [5] for curves of genus 1 and Weil [15] for arbitrary genus. Although the detailed approach may differ on surface, the basic idea is the same, which can be summarized as finding an idele class character η of the field F of rational functions on the underlying curve C of genus g over k such that

(i) the points occurring in the sum are precisely (the lifting of) the places of F of degree 1 at which η is unramified;

(ii) the term indexed by x in the character sum to be estimated is the value of η at a uniformizer at the place corresponding to x.

Denote by $\mathcal{D}$ the conductor of η, which is an effective divisor of F. The L-function $L(z,\eta)$ attached to a nonprincipal η is a polynomial in z of degree M equal to $2g-2$ plus the degree of $\mathcal{D}$. (Except for Bombieri's result, the curve C is P^1 and genus g is 0.) The character sum in question can then be expressed as the sum of the reciprocal roots of $L(z,\eta)$, and hence is bounded by $M\sqrt{q}$ by the Riemann hypothesis. In some situations, the bound M can be further improved, using the technique by Serre [10] and extended by Moreno and Moreno [9]. We shall not pursue this direction here.

The estimate of sums of the first kind over Galois rings can be used to construct families of cyclically different periodic sequences with low cross-correlation. This was done by Kumar et al. in [7], which we briefly recall in §3. An application of character sum estimates to the construction of codes with good parameters is also proposed there.

§2 Character sums over p-adic fields

Fix a p-adic local field K with residue field k of cardinality q, as in Introduction. By Hensel's lemma, elements in an algebraic closure of k can be uniquely lifted to zero and roots of unity with order prime to p in an algebraic closure of K, called Teichmuller elements. This enables us to lift polynomials over k to polynomials over $\mathcal{O}$ by lifting the roots and the leading coefficient of a polynomial. At a place v of $k(x)$ not equal to ∞, denote by π_v the monic irreducible polynomial over k whose roots constitute the points in v; at ∞ let π_∞ denote $1/x$.

Theorem 1. *Given a positive integer r, let ψ be an additive character of K trivial on $p^r\mathcal{O}$ but not on $p^{r-1}\mathcal{O}$. Let f, g be polynomials over $\mathcal{O}$ such that g is a lifting from a nonzero polynomial $\bar{g}$ over k, and $\bar{f} = f \pmod{p\mathcal{O}}$ is nonzero. Then there is a unique idele class character $\eta = \eta_{\psi,f/g}$ of $k(x)$ such that*

(1.a) η is unramified outside the poles of $\overline{f/g}$;

(1.b) The component η_v of η at a place v other than ∞ and poles of $\overline{f/g}$ is trivial on units and satisfies

$$\eta_v(\pi_v) = \psi(f/g(\gamma_1) + \cdots + f/g(\gamma_m)),$$

where m is the degree of π_v and $\gamma_1, ..., \gamma_m$ are the Teichmuller lifting of the roots of π_v.

(1.c) $\eta_\infty(\pi_\infty) = 1$.

If η is nonprincipal, then the degree D of its conductor is at least 2. In this case we have the estimate

$$|\sum_x \psi(f/g(x))| \leq (D-2)\sqrt{q},$$

where x runs through all points in $\mathcal{C}(K)$ which are not poles of f/g.

The conductor of η was explicitly described in [8], here we exhibit a few cases.

Example 1. $g = 1$. Suppose $f \not\equiv 0 \pmod{p^r}$, write it in base p expansion:

$$f \equiv f_0 + pf_1 + \cdots + p^{r-1}f_{r-1} \pmod{p^r},$$

where $f_0, ..., f_{r-1}$ are polynomials over $\mathcal{O}$ with coefficients being Teichmuller elements in K. Denote by d_i the degree of f_i. Suppose none of the nonzero d_i is divisible by p. (If f does not satisfy this condition, we may replace f by another polynomial satisfying the condition without changing the attached idele class character.) The weighted r-degree of f is defined as

$$w_r(f) = \max\{p^{r-1}d_0, p^{r-2}d_1, ..., d_{r-1}\}.$$

Then $\eta_{\psi,f}$ has conductor equal to $(w_r(f)+1)\infty$ so that $D = w_r(f) + 1$. As a consequence of this, we obtain a parametrization of the idele class characters of $k(x)$ ramified only at ∞:

Corollary 1. *Given a positive integer D, let r be the integer such that $p^{r-1} \leq D < p^r$. Let ψ be an additive character of K as in Theorem 1. Then $\eta_{\psi,f}$ parametrizes the idele class characters of $k(x)$, trivial at π_∞, with conductor $M\infty$ for $M \leq D+1$, as f runs through the polynomials of the form*

$$\sum_{0\leq t\leq r-1} p^t \sum_{m\geq 1, p\nmid m, p^{r-1-t}m\leq D} c_{t,m}x^m,$$

where $c_{t,m}$ are Teichmuller elements in K.

Example 2. Let f be as in Example 1. Then the idele class character $\eta_{\psi,f(1/x)}$ ramifies only at 0 with conductor equal to $(w_r(f)+1)0$. Hence $D = w_r(f)+1$ and we have a result similar to Corollary 1 with 0 replacing ∞ and $1/x$ replacing x.

Example 3. Let v be a place of $k(x)$ of degree 1 other than 0 and ∞. Let α be the Teichmuller lifting in K of the point in v. For f as in Example 1, the idele class character $\eta_{\psi,f(1/(x-\alpha))}$ ramifies only at v with conductor equal to $(w_r(f(1/(x-\alpha)))+1)v$, where

$$w_r(f(1/(x-\alpha))) = \max\{p^{r-1-i}(d_i+1) - 1 : 0 \leq i \leq r-1\}.$$

Example 4. Let v be a place of $k(x)$ of degree $m > 1$. Let $\alpha_1, ..., \alpha_m$ be the Teichmuller lifting of the points in v. Suppose g is a power of the lifting of π_v and the degree of f is less than the degree of g. Using partial fractions we can write f/g as $h_1(1/(x-\alpha_1)) + \cdots + h_m(1/(x-\alpha_m))$, where h_i's are polynomials with coefficients in the ring of integers of degree m unramified extension K_m of K and they are conjugate under the action of the Galois group of K_m over K. The idele class character $\eta_{\psi,f/g}$ is ramified only at v with conductor equal to $(w_r(h(1/(x-\alpha_1))) + 1)v$ and $D = (w_r(h(1/(x-\alpha_1))) + 1)m$.

We have the following analogus of Corollary 1.

Corollary 2. *Given a positive integer D, let r be the integer such that $p^{r-1} \leq D < p^r$. Let ψ be an additive character of K as in Theorem 1. Let v be a place of $k(x)$ other than ∞ and 0 and let g be the lifting of π_v. Then $\eta_{\psi,f}$ parametrizes the idele class characters of $k(x)$, trivial at π_∞, with order a power of p and conductor Mv for $M \leq D+1$ as f runs through the rational functions of the form*

$$\sum_{0\leq t\leq r-1} p^t \sum_{m\geq 1, p\nmid m, p^{r-1-t}m\leq D} (c_{t,m}/g)^{p^{r-1-t}m},$$

where $c_{t,m}$ runs through polynomials with coefficients being the Teichmuller elements in K and degree less than the degree of g.

For general f/g, decompose it into a sum of functions discussed in Examples 1-4 using partial fractions, the attached idele class character is the product of the character attached to each component, and its conductor is the sum of the conductor of the character arising from each component.

A similar result holds for a multiplicative character of K.

Theorem 2. *Given a positive integer s, let χ be a character of $K^\times$ which is trivial on $1+p^s\mathcal{O}$ but not trivial on $1+p^{s-1}\mathcal{O}$. Given polynomials f and g as in Theorem 1, there exists a unique idele class character $\omega = \omega_{\chi,f/g}$ of $k(x)$ such that*

(2.a) ω is unramified outside zeroes of $\overline{fg}$;

(2.b) The component ω_v of ω at a place v other than ∞ and zeroes of $\overline{fg}$ is trivial on units and satisfies

$$\omega_v(\pi_v) = \chi(f/g(\gamma_1)\cdots f/g(\gamma_m)),$$

where m is the degree of v and $\gamma_1, ..., \gamma_m$ are the Teichmuller lifting of the roots of π_v;

(2.c) $\omega_\infty(\pi_\infty) = 1$.

The conductor of ω can also be determined. As it is more complicated, we shall not present it here. The reader is referred to [8] for details. Suppose ω is nonprincipal with conductor of degree D, then it gives rise to the following estimate

$$|\sum_x \chi(f/g(x))| \leq (D-2)\sqrt{q},$$

where x runs through all points x on $C(K)$ which are not zeroes of fg.

An immediate consequence of Theorems 1 and 2 is

Theorem 3. *Let r, s, ψ, χ be as in Theorems 1 and 2. Let f_1, f_2, g_1, g_2 be polynomials over $\mathcal{O}$ such that g_1 and g_2 are liftings from nonzero polynomials over k, and $f_1 f_2 \pmod{p\mathcal{O}}$ is nonzero. Then $\xi = \eta_{\psi, f_1/g_1}\omega_{\chi, f_2/g_2}$ is the unique idele class character of $k(x)$ satisfying*

(3.a) ξ is unramified outside the poles of $\overline{f_1/g_1}$ and zeroes of $\overline{f_2 g_2}$;

(3.b) The component ξ_v of ξ at a place v other than ∞, poles of $\overline{f_1/g_1}$, and zeroes of $\overline{f_2 g_2}$ is trivial on units and satisfies

$$\xi_v(\pi_v) = \psi(f_1/g_1(\gamma_1) + \cdots + f_1/g_1(\gamma_m))\chi(f_2/g_2(\gamma_1)\cdots f_2/g_2(\gamma_m)),$$

where m is the degree of v and $\gamma_1, ..., \gamma_m$ are the Teichmuller lifting of the roots of π_v;

(3.c) $\xi_\infty(\pi_\infty) = 1$.

Again, if ξ is nonprincipal with conductor of degree D, then it gives the following estimate:

$$|\sum_x \psi(f_1/g_1(x))\chi(f_2/g_2(x))| \leq (D-2)\sqrt{q},$$

where x runs through all points x on $C(K)$ away from poles of f_1/g_1 and zeroes of $g_2 f_2$.

§3 Applications

The idele class characters of $k(x)$ attached to rational functions over $\mathcal{O}$ described above can be used to construct families of cyclically different periodic sequences with low cross-correlation. This was done in [7], where polynomials were used. We briefly describe their construction.

The nonzero Teichmuller elements in K form a cyclic group of order $q-1$; fix a generator β of the group. Given a positive integer D, let r be the integer such that $p^{r-1} \leq D < p^r$ and let ψ be an additive character of K as in Theorem 1. The idele class characters we shall use are $\eta_{\psi,f}$, parametrized by polynomials f as described in Corollary 1. To $\eta_{\psi,f}$ we attach the sequence $\vec{f}$ with its ith entry being $\psi(f(\beta^i))$ so that $\vec{f}$ is a sequence with period $L = q-1$. Note that the sequence obtained from shifting $\vec{f}$ forward by j coordinates is the sequence $\vec{g}$ where $g(x) = f(\beta^j x)$. Let $\mathcal{F}$ be a collection of sequences $\vec{f}$ obtained from choosing among nonzero polynomials f described in Corollary 1 such that $f(x) \neq g(\beta^i x)$ for $1 \leq i < L$ whenever f and g are chosen (including the case $f = g$). This gives a family of cyclically different sequences with values in p^r-th roots of unity, (in other words, the alphabet size is p^r), each has period $L = q-1$, and the correlation is dominated by $(D-1)\sqrt{q}+1$. The maximum size of $\mathcal{F}$ is $1/L$ times the size of the set S consisting of polynomials f in Corollary 1 satisfying $f(x) \neq f(\beta^i x)$ for $1 \leq i < L$. The size of S is at most $q^D - 1$, the number of nonzero polynomial parameters. However, if the smallest prime factor ℓ of L is $\leq D$, then $f(x) = p^{r-1}x^\ell$ is one of the parameters which do not lie in S since $f(x) = f(\beta^{L/\ell}x)$, and hence the size of $\mathcal{F}$ is less than $(q^D-1)/L$.

This point was overlooked in [7]. Those parameters with nonzero linear term are contained in S, which gives the lower bound $|S| \geq q^D - q^{(D-r)}$, which in turn shows that the size of $\mathcal{F}$ is at least $(q^D - q^{(D-r)})/L$.

Of course characters ramified at other places could be used to give a similar construction, but the size of the resulting family won't be as large in general.

Character sum estimates can also be used to construct codes with good parameters, as explained below.

The alphabet set A is a finite abelian group. We are interested in finding linear codes C over A of length n with large minimal distance. This means that the codewords in C should be as uniformly spread out as possible. For this purpose, we measure the distribution of the alphabets occurring in a codeword $v = (v_1, ..., v_n)$ of C. For $x \in A$, let $\Phi_v(x)$ denote the number of components of v equal to x. So Φ_v is a function on the abelian group A, its Fourier transform $\hat{\Phi}_v$ is a function on the group of characters $\hat{A}$ of A defined by

$$\hat{\Phi}_v(\psi) = \sum_{x \in A} \Phi_v(x)\psi(x) = \sum_{1 \leq i \leq n} \psi(v_i), \qquad \psi \in \hat{A}.$$

By Fourier inversion, we can recover the original function

$$\Phi_v(x) = \frac{1}{|A|} \sum_{\psi \in \hat{A}} \hat{\Phi}_v(\psi)\bar{\psi}(x) = \frac{n}{|A|} + \frac{1}{|A|} \sum_{\psi \in \hat{A}, \psi \neq 1} \hat{\Phi}_v(\psi)\bar{\psi}(x).$$

If for $\psi \neq 1$ there is a positive constant $\Lambda_v << n$ such that $|\hat{\Phi}_v(\psi)| \leq \Lambda_v$, then we get an estimate of the distribution of alphabets in v:

$$|\Phi_v(x) - \frac{n}{|A|}| \leq \frac{|A| - 1}{|A|} \Lambda_v.$$

Moreover, if A is a cyclic group $\mathbb{Z}_{|A|}$, then using Vinogradov inequality [12], one gets an estimate

$$|\sum_{h+1 \leq i \leq h+b} \Phi_v(i) - \frac{bn}{|A|}| \leq \Lambda_v \log |A|$$

for any $1 \leq b \leq |A|$ and any integer h.

Of course, the smaller Λ_v is, the more uniformly distributed the components of v are. It would be even better if there is a good bound Λ independent of the nonzero codewords v in C. This is where the character sum estimates in §2 come into play. Let S be a subset of the polynomials listed in Corollary 1 which is closed under multiplication by the Teichmuller elements in K and also closed under addition modulo p^r. To S we associate

$$C(S) = \{(\mathrm{Tr}(f(0)), \mathrm{Tr}(f(\beta)), ..., \mathrm{Tr}(f(\beta^{q-1}))) : f \in S\},$$

where Tr denotes the trace from K to $\mathbb{Q}_p$ followed by the natural map from $\mathcal{O}$ to $\mathcal{O}/p^r\mathcal{O}$. The condition on S implies that $C(S)$ is a linear code of length q with alphabet set $A = \mathbb{Z}/p^r\mathbb{Z}$. A codeword in $C(S)$ is called generic if it comes from a polynomial $f \not\equiv 0 \pmod{p}$, called a generic polynomial. Then the nontrivial Fourier coefficients of Φ_v for a generic codeword v in $C(S)$ is bounded by $\Lambda \leq (D-1)\sqrt{q}$. A nongeneric codeword w comes from a generic polynomial multiplied by p^s for some $s \geq 1$, and its components lie in p^sA. The Fourier coefficients of Φ_w with respect to the nontrivial characters of p^sA are also bounded by Λ.

This kind of code is very good when q is large in comparison to D. In particular, let $p = D = 2$, and S_2 consist of $f = (a + 2b)x$ with a, b Teichmuller elements in K. Note that when the first component of each codeword in $C(S_2)$ is repleced by parity-check, the resulting code is the $\mathbb{Z}_4$ Kerdock code $\mathcal{K}$ of length q described in [4] if q is an odd power of 2 and $q \geq 8$. It was computed by Calderbank and McGuire [2] that a nongeneric nonzero codeword in $\mathcal{K}$ has half of the entries 0 and the other half 2; while a generic codeword v satisfies

$$\Phi_v(i) = q/4 \pm \sqrt{q/8} \qquad \textit{for} \quad i \in \mathbb{Z}_4$$

and $\Phi_v(0) + \Phi_v(2) = q/2 = \Phi_v(1) + \Phi_v(3)$. This shows that the codewords in $\mathcal{K}$ are very uniformly distributed.

A main contribution of [4] is to point out that the known good nonlinear binary codes, such as Kerdock, Delsarte-Goethals, and Preparata codes, are the images of certain $\mathbb{Z}_4$-linear codes under the Gray map from $\mathbb{Z}_4$ to $(\mathbb{Z}_2)^2$ defined by $0 \mapsto (0,0), 1 \mapsto (1,0), 2 \mapsto (1,1)$, and $3 \mapsto (0,1)$, which is an isometry from Lee weight on $\mathbb{Z}_4$ to Hamming weight on $(\mathbb{Z}_2)^2$. Further, these $\mathbb{Z}_4$ codes are liftings of certain linear binary codes and the minimal Lee weight of the $\mathbb{Z}_4$ liftings is greater than the minimal Hamming weight of the initial binary codes. As the Gray map is an isometry, the binary codes obtained by applying the Gray map to the liftings have the same transmission rate as, but greater minimal distance than, the initial codes, and hence are better. This procedure was applied to other cyclic codes and indeed some improvements were obtained (cf. [3]). Yet this method seems to have a limited success so far. For instance, I am not aware of a good binary code constructed from a $\mathbb{Z}_8$-code. The lack of an isometry from $\mathbb{Z}_8$ with Lee weight to $(\mathbb{Z}_2)^3$ with Hamming weight seems to be regarded as an obstruction. I would like to re-examine this method from another angle.

As already mentioned, the Kerdock code is a very uniformly distributed $\mathbb{Z}_4$-code. In fact, the generic codewords have large Hamming weights under ANY bijection from $\mathbb{Z}_4$ to $(\mathbb{Z}_2)^2$, while the Hamming weight of the image of a nongeneric codeword depends very much on the bijection, and it has the highest weight if 2 is mapped to (1, 1) (assuming 0 mapped to $(0,0)$. This explains why the Gray map is chosen in order to give the binary Kerdock code the largest possible minimal distance $q - \sqrt{q/2}$. The $\mathbb{Z}_4$-Preparata code $\mathcal{P}$ is dual to $\mathcal{K}$. A minimally supported nonzero generic codeword has 4 components equal to 2, while a minimally supported generic codeword has 4 odd components and one component equal to 2. It is again clear that the Gray map gives the binary Preparata code the largest minimal distance 6.

Let S_4 consist of $f = (a + 2b + 4c)x$ with a, b, c Teichmuller elements in K. Replacing the first component of each codeword in $C(S_4)$ by parity-check, we get the $\mathbb{Z}_8$ Kerdock code. It is easy to check that for $q > 98$ any map from $\mathbb{Z}_8$ to $(\mathbb{Z}_2)^3$ sending 0 to $(0,0,0)$ so that the image of 4 has Hamming weight 3, and the images of 2 and 6 have Hamming weight 2 will map the $\mathbb{Z}_8$ Kerdock code to a binary code with the same transmission rate $\frac{\log_2 q}{q}$ as before, but with minimal distance greater than q. On the other hand, if we use S_4' consisting of $f = ax + 2bx + 4(cx + dx^3)$ with a, b, c, d Teichmuller elements in K, then the same construction yields, for $q > 144$, binary codes with transmission rate $\frac{4 \log_2 q}{3q}$ and minimal distance greater than q. Both $\mathbb{Z}_8$ codes modulo 4 reduce to $\mathcal{K}$. Therefore there is no reason to stop at the $\mathbb{Z}_4$ codes!

Philosophically speaking, the $\mathbb{Z}_4$-Kerdock code is good because it is uniformly distributed; I have no explanation for the $\mathbb{Z}_4$-Preparata code except that it is dual to a good code. It would be desirable to show that the dual of a good code is again good. As discussed above, we obtain binary codes with at least the same transmission rate and improved minimal distance as we lift the initial code to codes modulo higher and higher power of 2. It would be interesting to find the best minimal distance and transmission rate of binary codes obtained using this procedure.

Observe that $C(S)$ is actually an algebraic geometry (AG) code of the projective line over the ring $\mathcal{O}/p^r\mathcal{O}$. As is well-known, using modular curves, Tsfasman, Vladut and Zink [11] found AG codes better than the Gilbert-Varshamov bound. In light of the above discussion, one might hope to get even better nonlinear codes by considering AG codes of curves over $\mathcal{O}/p^r\mathcal{O}$. Naturally the first step would be to extend the character sum estimates considered in §2 to curves of arbitrary genus. The case $r = 1$ is the same as over a finite field, which was done by Bombieri [1] for sums involving an additive character. Walker's thesis [14] studied AG codes over rings in general, and her joint work with Voloch [13] estimated some character sums arising from an elliptic curve over $\mathcal{O}/p^r\mathcal{O}$.

References

[1] E. Bombieri: On exponential sums in finite fields, Amer. J. Math. 88 (1966), 71-105.

[2] A. R. Calderbank and G. McGuire: $\mathbb{Z}_4$-linear codes obtained as projections of Kerdock and Delsart-Goethals codes, Lin. Alg. Appl., vol. 226-228 (1995), 647-665.

[3] A. R. Calderbank, G. McGuire, P. V. Kumar, and T. Helleseth: Cyclic codes over $\mathbb{Z}_4$, locator polynomials and Newton's identities, IEEE Trans. Inform. Theory, vol. 42 (1966), 217-226.

[4] A. R. Hammons, Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé: The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals and related codes, IEEE Trans. Inform. Theory, vol. 40, 301-319, 1994.

[5] H. Hasse: Beweis des Analogons der Riemannschen Vermutung für die Artinschen und F. K. Schmidtschen Kongruenzzetafunktionen in gewissen elliptischen Fällen. Ges. d. Wiss. Nachrichten. Math. Phys. Klasse, 1933, Heft 3, 253-262.

[6] T. Helleseth, V. Kumar and A. Shanbhag: Exponential sums over Galois rings and their applications. In: *Finite Fields and Applications*, S. Cohen and H.

Niederreiter edited, London Math. Soc. Lecture Note Series 233, Cambridge Univ. Press (1996), 109-128.

[7] P. V. Kumar, T. Helleseth, and A. R. Calderbank: An upper bound for some exponential sums over Galois rings and applications, IEEE Transactions on Information Theory 41 (1995), 456-468.

[8] W.-C. W. Li: Character sums over p-adic fields, submitted.

[9] O. Moreno and C. Moreno: A p-adic Serre bound, Finite Fields and Their Applications, to appear.

[10] J-P. Serre: Sur le nombre des points rationells d'une courbe algebrique sur un corps fini, C. R. Acad. Sci. Paris 296 (1983), 397-402.

[11] M. A. Tsfasman, S. G. Vladut, and Th. Zink: Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound, Math. Nachr. 109 (1982), 21-28.

[12]. M. Vinogradov: Sur la distribution des résidues des puissances, Ž. Fiz.-Mat. Obŝĉ. Permsk. Gos. Univ., no. 1 (1918), 94-98.

[13] J. Voloch and J. Walker: Euclidean weights of codes from elliptic curves over rings, preprint (1997).

[14] J. Walker: Algebraic geometric codes over rings, submitted.

[15] A. Weil: Sur les Courbes Algébriques et les Variétés Qui s'en Deduisent, Hermann, Paris (1948).

[16] A. Weil: On some exponential sums, Proc. National Academy of Science 34 (1948), 204-207.

Explicit modular towers

Noam D. Elkies
Department of Mathematics, Harvard University
Cambridge, MA 02138 (e-mail: `elkies@math.harvard.edu`)

Abstract. We give a general recipe for explicitly constructing asymptotically optimal towers of modular curves such as $\{X_0(l^n)\}_{n>1}$. We illustrate the method by giving equations for eight towers with various geometric features. We conclude by observing that such towers are all of a specific recursive form and speculate that perhaps every tower of this form which attains the Drinfeld-Vlăduţ bound is modular.

Introduction. Explicit equations for modular curves have attracted interest at least since the classical work of Fricke and Klein. Recent renewed interest in such equations has been stimulated on the one hand by the availability of software for symbolic computation and on the other hand by specific applications. In [E] we considered the use of modular curves to count rational points on elliptic curves over large finite fields, and illustrated some other applications of equations for the curves $X_0(N)$ with small N (say $N < 10^3$). Another kind of application is to coding theory: good Goppa codes [G] require curves of large genus g over a fixed finite field $k = \mathbf{F}_q$ whose number of rational points grows as a positive multiple of g. Drinfeld and Vlăduţ showed that as $g \to \infty$ no multiple greater than $(q^{1/2} - 1)g$ is possible. Ihara [I] and, independently, Tsfasman, Vlăduţ, and Zink [TVZ] showed that this upper bound is attained by the supersingular points on appropriate modular curves when q is a square. For this application modular curves — elliptic, Shimura, or Drinfeld — are needed whose level is too high to apply the methods of [E] directly, and in general one does not expect to have any pleasant model for a curve of high genus. However, if the curve is of smooth level then it tops a tower of $O(\log q)$ covers of low degree, and one may hope to obtain equations for the curve by writing those covers explicitly.

In this paper we show how to do this recursively for towers such as $\{X_0(l^n)\}_{n>1}$. It turns out that only information about the first few levels of the tower is needed, and that this information can be obtained for modular elliptic curves using the methods of [E], and for some Shimura curves using only the ramification structure. We then illustrate the method by giving explicit formulas for eight asymptotically optimal towers: six of elliptic modular curves, namely $X_0(l^n)$ for $l = 2, 3, 4, 5, 6$, and $X_0(3 \cdot 2^n)$; and two of Shimura modular curves. Over any finite field whose characteristic does not divide the level of these modular curves, the towers are tamely ramified, making it easy to calculate the genus of every curve in the tower. [This contrasts with the wildly ramified tower of [GS1], whose genus computation required some ingenuity; we show elsewhere that that tower too is modular, of Drinfeld type.] For each finite field k over which one of our towers is asymptotically optimal, the optimality can then be shown by elementary means, independent of the tower's modular provenance, by exhibiting the coordinates of the rational (supersingular) points. These formulas may also have other uses, e.g. in finding explicit modular parametrizations of elliptic curves with smooth conductor, or in connection with generalizations of the arithmetic-geometric mean (which corresponds to the $X_0(2^n)$ tower) as in [S1,S2]; we hope to pursue these connections in future papers. We conclude this paper with a speculation concerning the modularity of "any" asymptotically optimal tower.

The curves $X_0(l^n)$. Fix a prime $l > 1$. For positive n, the elliptic modular curve $X_0(l^n)$ over any field k in which $l \neq 0$ parametrizes elliptic curves with a cyclic l^n-isogeny, or equivalently sequences of l-isogenies

$$E_0 \to E_1 \to E_2 \to \cdots \to E_n \tag{1}$$

such that the composite isogeny $E_{j-1} \to E_{j+1}$ of degree l^2 is cyclic for each j with $0 < j < n$. Thus for each $m = 0, 1, \ldots, n$ there are $n + 1 - m$ maps $\pi_j : X_0(l^n) \to X_0(l^m)$ obtained by extracting for some $j = 0, 1, \ldots, n - m$ the cyclic l^m-isogeny $E_j \to E_{j+m}$ from (1). Each of these maps has degree l^{n-m}, unless $m = 0$ when the degree is $(l + 1)l^{n-1}$. In particular we have a tower of maps

$$X_0(l^n) \xrightarrow{\pi_0} X_0(l^{n-1}) \xrightarrow{\pi_0} X_0(l^{n-2}) \xrightarrow{\pi_0} \cdots \xrightarrow{\pi_0} X_0(l^2) \xrightarrow{\pi_0} X_0(l), \tag{2}$$

each map being of degree l. Each $X_0(l^n)$ also has an Atkin-Lehner involution $w_l = w_l^{(n)}$, taking a cyclic l^n-isogeny to its dual isogeny, and the sequence (1) to the sequence

$$E_n \to \cdots \to E_2 \to E_1 \to E_0 \tag{3}$$

of dual isogenies. We thus have

$$w_l^{(m)} \circ \pi_j = \pi_{n-m-j} \circ w_l^{(n)}, \tag{4}$$

where π_j, π_{n-m-j} are our jth and $(n - m - j)$th maps from $X_0(l^n)$ to $X_0(l^m)$.

When $k = \mathbf{C}$, we may regard $X_0(N)$ as the quotient of the extended upper half-plane $\mathcal{H}^* = \mathcal{H} \cup \mathbf{P}^1(\mathbf{Q})$ by $\Gamma_0(N)$. Then the $\Gamma_0(l^n)$ orbit of a point $\tau \in \mathcal{H}$ parametrizes the isogeny between the elliptic curves with period lattices $\mathbf{Z} + \tau\mathbf{Z}$ and $l^{-n}\mathbf{Z} + \tau\mathbf{Z}$, the map π_j takes the $\Gamma_0(l^n)$ orbit of τ to the $\Gamma_0(l^m)$ orbit of $l^j\tau$, and the involution $w_l^{(n)}$ is represented by $\tau \longleftrightarrow -1/l^n\tau$.

Now the key observation is that explicit formulas for $X_0(l)$, $X_0(l^2)$, together with the involutions $w_l^{(1)}, w_l^{(2)}$ of these curves and the map $\pi_0 : X_0(l^2) \to X_0(l)$ between them, suffice to exhibit the entire tower (2) explicitly:

Proposition. *For $n \geqslant 2$ the product map*

$$\pi = \pi_0 \times \pi_1 \times \pi_2 \times \cdots \times \pi_{n-2} : X_0(l^n) \to (X_0(l^2))^{n-1} \tag{5}$$

is a 1:1 map from $X_0(l^n)$ to the set of $(P_1, P_2, \ldots, P_{n-1}) \in (X_0(l^2))^{n-1}$ such that

$$\pi_0(w_l^{(2)}(P_j)) = w_l^{(1)}(\pi_0(P_{j+1})) \tag{6}$$

for each $j = 1, 2, \ldots, n - 2$.

Informally speaking, we get from $X_0(l^2)$ up to $X_0(l^n)$ by iterating $n-2$ times the involution $w_l^{(2)}$ composed with the "l-valued involution" $\pi_0^{-1} w_l^{(1)} \pi_0$. Of course the maps $\pi_j : X_0(l^n) \to X_0(l^m)$ (for $m \geqslant 2$) are then simply

$$(P_1, \ldots, P_{n-1}) \mapsto (P_{j+1}, \ldots, P_{j+m-1}), \tag{7}$$

and the involution $w_l^{(n)}$ is

$$(P_1, P_2, \ldots, P_{n-2}, P_{n-1}) \longleftrightarrow (w_l^{(2)} P_{n-1}, w_l^{(2)} P_{n-2}, \ldots, w_l^{(2)} P_2, w_l^{(2)} P_1), \tag{8}$$

i.e. reversing the order of $P_1, \dots, P_{n-1}$ and applying $w_l^{(2)}$ to each coordinate.

Proof[1]: That the map is 1:1 to its image is clear, because a sequence (1) of l-isogenies is determined by the l^2-isogenies $E_{j-1} \to E_{j+1}$ parametrized by the jth coordinate of π $(0 < j < n)$. Now $(P_1, \dots, P_{n-1})$ is in the image of π if and only if the l^2-isogenies parametrized by $P_1, \dots, P_{n-1}$, regarded as sequences $E_0^j \to E_1^j \to E_2^j$ of l-isogenies, fit together to form a sequence (1) with $E_i^j = E_{i+j}$, i.e. if and only if the isogenies $E_1^j \to E_2^j$ and $E_0^{j+1} \to E_1^{j+1}$ coincide for each $j = 1, 2, \dots, n-2$. But these isogenies are represented by the points $\pi_1(P_j)$ and $\pi_0(P_{j+1})$ on $\mathrm{X}_0(l)$. Thus the necessary and sufficient condition is that

$$\pi_1(P_j) = \pi_0(P_{j+1}) \tag{9}$$

for each $j = 1, 2, \dots, n-2$; applying $w_l^{(1)}$ to both sides, and then (4) to $w_l^{(1)}(\pi_1(P_j))$, then yields the equivalent form (6). □

Examples: The cases $l = 2, 3, 5$. Our formulas are particularly simple when $\mathrm{X}_0(l^2)$ (and thus also $\mathrm{X}_0(l)$) has genus 0, for then we may use a Hauptmodul (or for that matter any rational parameter[2]) of $\mathrm{X}_0(l^2)$ to regard $P_1, \dots, P_{n-1}$ as $n-1$ rational coordinates on $\mathrm{X}_0(l^n)$, and (6) as the $n-2$ algebraic relations on those coordinates that determine the curve $\mathrm{X}_0(l^n)$. This happens for $l = 2, 3, 5$; we exhibit formulas for each of these cases.

In the first two cases the cover $\pi_0 : \mathrm{X}_0(l^2) \to \mathrm{X}_0(l)$ is cyclic.[3] For $l = 2$ we parametrize $\mathrm{X}_0(l^2) = \mathrm{X}_0(4)$ by

$$\xi(\tau) := 1 + \frac{1}{8}\left(\frac{\eta(\tau)}{\eta(4\tau)}\right)^8 = \frac{1}{8}(q^{-1} + 20q - 62q^3 + 216q^5 - 641q^7 + - \cdots). \tag{10}$$

where as usual $q = e^{2\pi i\tau}$ and η is the weight-$\frac{1}{2}$ modular form $\prod_{r=1}^{\infty}(1-q^r)$. Using the functional equation

$$\eta(-1/\tau) = (\tau/i)^{1/2}\eta(\tau), \tag{11}$$

we find that the involution $w_2^{(2)}$ takes $\xi(\tau)$ to

$$\xi(-1/4\tau) = 1 + 32\left(\frac{\eta(4\tau)}{\eta(\tau)}\right)^8 = 1 + \frac{4}{\xi(\tau)-1} = \frac{\xi(\tau)+3}{\xi(\tau)-1}. \tag{12}$$

Let h_2 be the $\mathrm{X}_0(2)$ Hauptmodul

$$h_2(\tau) = \left(\frac{\eta(\tau)}{\eta(2\tau)}\right)^{24} = q^{-1} - 24 + 276q - 2048q^2 + 11202q^3 - + \cdots. \tag{13}$$

---

[1] More properly, a proof sketch, since we suppress some details, such as what happens at the cusps. To show that our formulas extend to the cusps one may either quote general facts about maps between affine and projective algebraic curves, or regard the cusps as parametrizing isogenies between Tate curves. Also, two cyclic l-isogenies may determine the same point on $\mathrm{X}_0(l)$ without being isomorphic; the necessary and sufficient condition is that they become isomorphic over the algebraic closure.

[2] A "Hauptmodul" is a rational parameter with a pole of leading coefficient 1 at the infinite cusp, i.e. a degree-1 rational function of the form $q^{-1} + O(1)$.

[3] For any N, the cover $\pi_0 : \mathrm{X}_0(N^2) \to \mathrm{X}_0(N)$ is cyclic if and only if the unit group of $\mathbf{Z}/N\mathbf{Z}$ has exponent 2, which happens when $N|24$. When $N = 1, 2, 3, 4, 6$, it is furthermore true that ± 1 are the only units of $\mathbf{Z}/N\mathbf{Z}$, and then the cyclic N-isogenies $E_1 \to E_0, E_2$ together with the Weil pairing on $E_1[N]$ determine a complete level-N structure on E_1 mod ± 1, i.e. $\mathrm{X}_0(l^2) \cong \mathrm{X}(l)$.

Computing the map $\pi_0 : \mathrm{X}_0(4) \to \mathrm{X}_0(2)$ then amounts to writing h_2 as a rational function in ξ. We do this by in effect expanding this function as a continued fraction. Necessarily that function has degree 2 with a simple pole at the cusp $\xi = \infty$. But then $h_2 - 8\xi + 24$ is a rational function of degree 1 in ξ with a simple zero at ∞, i.e. the inverse of a polynomial of degree 1. Comparing the q-expansions of $1/(h_2 - 8\xi + 24)$ and ξ, we find that this polynomial is $(\xi + 1)/32$ and recover the formula

$$h_2(\tau) = 8\frac{(\xi(\tau)+1)^2}{(\xi(\tau)-1)}. \tag{14}$$

Using our formula (12) for the involution $w_2^{(2)}$ we then obtain also

$$h_2(\tau) = \frac{64}{\xi(-1/4\tau)^2 - 1}. \tag{15}$$

But $w_2^{(1)}$ acts on $\mathrm{X}_0(2)$ by $h_2 \leftrightarrow 2^{12}/h_2$ (again by (11)). Thus $h_2(2\tau)$ is both $64(\xi(\tau)^2 - 1)$ and $64/(\xi(-1/8\tau)^2 - 1)$. Equating these two expressions yields an equation for the modular curve $\mathrm{X}_0(8)$; more generally we now deduce from our Proposition the following explicit equations for the modular curve $\mathrm{X}_0(2^n)$ for each $n > 1$:

Let x_j $(0 < j < n)$ be the rational function $\xi(2^{j-1}\tau)$ on that curve (this is the coordinate P_j of the Proposition); then $(x_1, \ldots, x_{n-1})$ identifies $\mathrm{X}_0(2^n)$ with the curve in $(\mathbf{P}^1)^{n-1}$ specified by the $n-2$ equations

$$(x_j^2 - 1)(z_{j+1}^2 - 1) = 1 \qquad (j = 1, \ldots, n-2), \tag{16}$$

where

$$z_j := (x_j + 3)/(x_j - 1) \tag{17}$$

is obtained from x_j by the involution $w_2^{(2)}$.

Curiously we obtain analogous equations for $\mathrm{X}_0(3^n)$ by replacing the exponent 2 by 3 in (16) and, as if to compensate, changing the constant term 3 to 2 in (17): the curve $\mathrm{X}_0(3^n)$ is isomorphic with the locus of $(x_1, \ldots, x_{n-1})$ in $(\mathbf{P}^1)^{n-1}$ satisfying

$$(x_j^3 - 1)(z_{j+1}^3 - 1) = 1 \qquad (j = 1, \ldots, n-2). \tag{18}$$

where

$$z_j := (x_j + 2)/(x_j - 1). \tag{19}$$

Here the coordinate functions x_j on $\mathrm{X}_0(3^n)$ are $\xi(3^{j-1}\tau)$, where

$$\xi(\tau) = 1 + \frac{1}{3}\left(\frac{\eta(\tau)}{\eta(9\tau)}\right)^3 = \frac{1}{3}(q^{-1} + 5q - 7q^5 + 3q^8 + 15q^{11} - 32q^{14} \cdots), \tag{20}$$

so ξ generates the field of rational functions on $\mathrm{X}_0(9)$. The involution $w_3^{(2)}$ takes this ξ to $1 + 3/(\xi - 1) = (\xi + 2)/(\xi - 1)$, whence (19); the Hauptmodul

$$h_3(\tau) = \left(\frac{\eta(\tau)}{\eta(3\tau)}\right)^{12} = q^{-1} - 12 + 54q - 76q^2 - 243q^3 + 1188q^4 \cdots \tag{21}$$

goes to $3^6/h_3$ under $w_3^{(1)}$, and $h_3(3\tau)$ is both $27/(\xi(-1/27\tau)^3 - 1)$ and $27(\xi(\tau)^3 - 1)$, from which the equations (18,19) for $X_0(3^n)$ follow thanks to our Proposition.

Note that the already simple equations for $X_0(2^k)$, $X_0(3^k)$ simplify even further in characteristic 3, 2 respectively: taking $y_j = 1 - x_j^{-1}$ in (16,17) and setting $3 = 0$ yields

$$y_{j+1}^2 = y_j - y_j^2, \tag{22}$$

and the same substitution in (18,19) with $2 = 0$ produces

$$y_{j+1}^3 = y_j^3 + y_j^2 + y_j. \tag{23}$$

In this guise these asymptotically optimal towers were obtained by Garcia and Stichtenoth [GS2, Examples C,D], independent (as in [GS1]) of their modular interpretation. In both cases the supersingular points are the poles of y_1 and thus of all the y_j.

Finally for $l = 5$ we obtain

$$P(x_j)P(z_{j+1}) = 125 \qquad (j = 1, \ldots, n-2), \tag{24}$$

where

$$P(X) := X^5 + 5X^3 + 5X - 11, \qquad z_j := (x_j + 4)/(x_j - 1). \tag{25}$$

Here the coordinate x_j is $\xi(5^{j-1}\tau)$ where

$$\xi(\tau) = 1 + \frac{\eta(\tau)}{\eta(25\tau)} = q^{-1} - q + q^4 + q^6 - q^{11} - q^{14} + q^{21} + q^{24} - q^{26} \cdots. \tag{26}$$

As usual, z_j is the image of x_j under $w_l^{(2)}$, and $P(x_j) = h_5(5\tau)$ where h_5 is the $X_0(5)$ Hauptmodul

$$h_5(\tau) = \left(\frac{\eta(\tau)}{\eta(5\tau)}\right)^6 = q^{-1} - 6 + 9q + 10q^2 - 30q^3 + 6q^4 - 25q^5 \cdots \tag{27}$$

with $h_5(\tau)h_5(-1/5\tau) = 125$. The polynomial $P(X)$ is necessarily not as simple as the polynomials $X^2 - 1$, $X^3 - 1$ occurring in (16,18), because the cover $X_0(25) \to X_0(5)$ is not cyclic. It is, however, dihedral, as may be seen from the fact that $P(W - 1/W) = W^5 - 11 - W^{-5}$.

First variation: composite l. The assumption that l be prime was not necessary; the entire description carries over to the composite case, except for the incidental point that the degree of the maps $\pi_j : X_0(l^n) \to X(1)$ is given by a formula more complicated than $(l+1)l^{n-1}$ [namely $l^n \prod_{p|l}(1 + \frac{1}{p})$]. For instance we exhibit formulas for the cases $l = 4, 6$, where the cover $\pi_0 : X_0(l^2) \to X_0(l)$ is still cyclic.

In the first case $l = 4$ the curve $X_0(l^2) = X_0(16)$ is still rational, and we obtain formulas remarkably similar to those for $l = 2, 3$ by choosing

$$\xi(\tau) = 1 + \frac{1}{2}\frac{\eta^2(\tau)\eta(8\tau)}{\eta(2\tau)\eta^2(16\tau)} = \frac{1}{2}(q^{-1} + 2q^3 - q^7 - 2q^{11} + 3q^{15} + 2q^{19} \cdots) \tag{28}$$

as a rational coordinate on $X_0(16)$. Then $w_4^{(2)}$ takes ξ to $(\xi+1)/(\xi-1)$. The $X_0(4)$ Hauptmodul

$$h_4(\tau) = \left(\frac{\eta(\tau)}{\eta(4\tau)}\right)^8 = q^{-1} - 8 + 20q - 62q^3 + 216q^5 - 641q^7 + - \cdots \tag{29}$$

(cf. (10)) is mapped by $w_4^{(1)}$ to $4^4/h_4$, and we compute

$$h_4(4\tau) = \frac{16}{\xi(-1/64\tau)^4 - 1} = 16(\xi(\tau)^4 - 1). \tag{30}$$

Therefore $X_0(4^n)$ is isomorphic with the locus of $(x_1, \ldots, x_{n-1})$ in $(\mathbf{P}^1)^{n-1}$ satisfying

$$(x_j^4 - 1)(z_{j+1}^4 - 1) = 1 \qquad (j = 1, \ldots, n-2), \tag{31}$$

where

$$z_j := (x_j + 1)/(x_j - 1), \tag{32}$$

the coordinate functions x_j on $X_0(4^n)$ being $\xi(4^{j-1}\tau)$. Of course the resulting curves also occur in the $X_0(2^n)$ tower, but this fact is far from obvious from comparison of the formulas (16,17) and (31,32).

The case of $l = 6$ is slightly more complicated because the curve $X_0(l^2)$ is no longer rational. It is, however, an elliptic curve with a simple Weierstrass equation: the ring of rational functions on $X_0(36)$ regular except possibly at the cusp $\tau = i\infty$ is generated by

$$\begin{aligned} \xi(\tau) &= \frac{\eta(12\tau)\eta^3(18\tau)}{\eta(6\tau)\eta^3(36\tau)} = q^{-2} + q^2 + q^8 - q^{14} - q^{20} + q^{26} + 2q^{32}\cdots, &(33)\\ \gamma(\tau) &= \frac{\eta^4(12\tau)\eta^2(18\tau)}{\eta^2(6\tau)\eta^4(36\tau)} = q^{-3} + 2q^3 + q^9 - 2q^{15} - 2q^{21} + 2q^{27} + 4q^{33}\cdots, &(34) \end{aligned}$$

related by the Weierstrass equation

$$\gamma^2 = \xi^3 + 1. \tag{35}$$

The involution $w_6^{(2)}$ has a fixed point at $i/6$. An involution of an elliptic curve which has a fixed point must be multiplication by -1 composed with a translation. Thus to determine $w_6^{(2)}$ we need only find the image of one point. It is easiest to do this with the cusp $\tau = i\infty$: its image is the cusp $\tau = 0$, at which $(\xi, \gamma) = (2, 3)$ (a 6-torsion point on the curve (35)). It remains only to find the map from $X_0(36)$ to $X_0(6)$ and the involution $w_6^{(1)}$. We use the Hauptmodul

$$h_6(\tau) = \frac{\eta^5(\tau)\eta(3\tau)}{\eta(2\tau)\eta^5(6\tau)} = q^{-1} - 5 + 6q + 4q^2 - 3q^3 - 12q^4 - 8q^5 + 12q^6 \cdots. \tag{36}$$

Then $w_6^{(1)}$ takes h_6 to $72/h_6$, and by comparing q-expansions we find

$$h_6(6\tau) = \xi^3(\tau) - 8. \tag{37}$$

We thus identify $X_0(6^n)$ with the curve of $(n-1)$-tuples $((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}))$ of points on the elliptic curve $y^2 = x^3 + 1$ satisfying the $n-2$ conditions

$$(x_j^3 - 8)(z_{j+1}^3 - 8) = 72 \qquad (j = 1, \ldots, n-2), \tag{38}$$

where

$$z_j := \left(\frac{y_j+3}{x_j-2}\right)^2 - x_j - 2 \tag{39}$$

is the x-coordinate of the point $(2, 3) - (x, y)$ on $y^2 = x^3 + 1$. Unlike the curves in the $\mathrm{X}_0(4^n)$ tower, these curves $\mathrm{X}_0(6^n)$ are new; to be sure they could also be exhibited as composita of the already known covers $\mathrm{X}_0(2^n)/\mathrm{X}(1)$ and $\mathrm{X}_0(3^n)/\mathrm{X}(1)$, but those models are much harder to work with because of the complicated singularities above the branch points $j = 0, 12^3, \infty$.

Second variation: changing the base of the tower. Instead of the tower of modular curves $\mathrm{X}_0(l^n) = \mathcal{H}^*/\Gamma_0(l^n)$ we could use $\mathcal{H}^*/(\Delta \cap \Gamma_0(l^n))$ where Δ is some other congruence subgroup of $\mathrm{PGL}_2(\mathbf{Q})$, as long as the modulus of the congruence is prime to l. For instance, given $N > 1$ with $(l, N) = 1$ we could use the tower $\mathrm{X}_0(Nl^n)$ of curves parametrizing sequences of l-isogenies between pairs of elliptic curves related by a cyclic N-isogeny. Again these curves with $n > 1$ form a tower related by maps π_j of l-power degree and admitting involutions $w_l^{(n)}$, and knowing these maps and involutions for $n = 1, 2$ yields explicit formulas for $\mathrm{X}_0(Nl^n)$ for all $n > 1$ as in our Proposition.[4]

We illustrate with the case $l = 2$, $N = 3$. In this case the first two curves $\mathrm{X}_0(6), \mathrm{X}_0(12)$ are rational and we can mimic our procedure for the towers $\mathrm{X}_0(l^n)$ with $l = 2, 3, 4$. Our $(n-1)$ coordinates on $\mathrm{X}_0(3 \cdot 2^n)$ will be $x_j = \xi(2^{j-1}\tau)$ $(0 < j < n)$ where

$$\xi(\tau) = \frac{\eta^4(4\tau)\eta^2(6\tau)}{\eta^2(2\tau)\eta^4(12\tau)} = q^{-1} + 2q + q^3 - 2q^5 - 2q^7 + 2q^9 + 4q^{11} \cdots \tag{40}$$

(cf. (34)) is a Hauptmodul for $\mathrm{X}_0(12)$. It is this time more convenient to let h_6 be the $\mathrm{X}_0(6)$ Hauptmodul

$$h_6(\tau) = \left(\frac{\eta(2\tau)\eta^3(3\tau)}{\eta(\tau)\eta^3(6\tau)}\right)^3 = q^{-1} + 3 + 6q + 4q^2 - 3q^3 - 12q^4 - 8q^5 \cdots \tag{41}$$

which differs by 8 from our choice in (36). We may then represent $w_2^{(1)}$ and $w_2^{(2)}$ by $\tau \longleftrightarrow (2\tau - 1)/(6\tau - 2)$ and $\tau \longleftrightarrow (4\tau + 3)/(4\tau + 4)$; these involutions take h_6 to $-8/h_6$ and ξ to $(3 - \xi)/(1 + \xi)$. By computing the quadratic map $\mathrm{X}_0(12) \to \mathrm{X}_0(6)$ we find that this time

$$h_6(2\tau) = \frac{-8}{\xi(w_2^{(2)}\tau)^2 - 1} = \xi(\tau)^2 - 1. \tag{42}$$

Thus the equations on $x_1, \ldots, x_{n-1}$ defining $\mathrm{X}_0(3 \cdot 2^n)$ are

$$(x_j^2 - 1)(z_{j+1}^2 - 1) = -8 \qquad (j = 1, \ldots, n-2), \tag{43}$$

where

$$z_j := (3 - x_j)/(1 + x_j). \tag{44}$$

---

[4] As with $\mathrm{X}_0(6^n)$ we could also obtain $\mathrm{X}_0(Nl^n)$ as a compositum of $\mathrm{X}_0(N)$ and $\mathrm{X}_0(l^n)$, but the resulting model is highly singular. Warning: on $\mathrm{X}_0(Nl^n)/\mathbf{C}$ the involution $w_l^{(n)}$ is given not by $\tau \leftrightarrow -1/l^n\tau$ but by a fractional linear transformation of the same determinant that reduces mod N to an element of $\Gamma_0(N)$. We do still have a simple formula $\tau \leftrightarrow -1/Nl^n\tau$ for the product of $w_l^{(n)}$ with the Atkin-Lehner involution w_N.

Note that in this case the curves in our tower also have an involution w_3 commuting with all the $w_2^{(n)}$; we find that this involution is $x_j \leftrightarrow -3/x_j$. That this in fact acts on our model of $X_0(3 \cdot 2^n)$ is easy to check after writing (43,44) in the equivalent form

$$(x_{j+1} - 1)x_j^2 = x_{j+1}^2 + 3x_{j+1}. \tag{45}$$

Third variation: Shimura modular curves. Shimura curves generalize the classical elliptic modular curves: instead of $\mathcal{H}^*/\Gamma$ for an arithmetic subgroup Γ of $\mathrm{PGL}_2(\mathbf{Q})$, they are the quotients $\mathcal{H}/\Gamma$ by an arithmetic subgroup of a quaternion algebra A over some totally real number field K, with A ramified at all but one of the infinite places of K. Instead of elliptic curves, these Shimura curves parametrize principally polarized abelian varieties with endomorphisms by A and some extra structure determined by the choice of Γ. There are Shimura curves $\mathcal{X}_0(I)$ (I an ideal of K coprime with the discriminant of A) analogous to $X_0(N)$, which have Atkin-Lehner involutions and form towers, and whose reductions at a prime of K are asymptotically optimal over the quadratic extensions of its residue field. These towers may be obtained from their first two levels by the recipe of our Propositions.

Unlike the classical $X_0(N)$, the analogous Shimura curves $\mathcal{X}_0(I)$ have no cusps. Thus the curves and maps between them cannot be computed using q-expansions. Even worse, in general we do not even have explicit equations for the abelian varieties parametrized by these curves. Nevertheless we can in many cases use the ramification behavior of the covers to determine the necessary maps completely. We illustrate this with two examples which have the additional feature of involving only cyclic covers which become unramified after finitely many steps and thus also occur in class-field towers.[5]

We start with K, A such that A^* contains an arithmetic subgroup Δ which is also a triangle group. Such Δ have been classified completely [T]: there are 76, in 18 quaternion algebras (not including the nine triangle subgroups of $\mathrm{PGL}_2(\mathbf{Q})$ with one or more cusps among the vertices). For our first example we take $K = \mathbf{Q}(\sqrt{3}\,)$ and $A/K =$ the quaternion algebra ramified at $(\sqrt{3}\,)$ and at one infinite place, and choose for Δ the group called $\Gamma^{(+)}(A, O_1) = \Gamma^{(*)}(A, O_1)$ in [T], which is identified there with the $(2, 4, 12)$ triangle group. We shall construct the tower $\{\mathcal{X}_0(\wp_2^n)\}_{n>1}$, where $\wp_2$ is the prime of K of residue field $\mathbf{F}_2$.

The curve $\mathcal{X}(1) = \mathcal{H}/\Delta$ is rational. We choose a coordinate J taking the values $1, 0, \infty$ at the elliptic points of order $2, 4, 12$. The curve $\mathcal{X}_0(\wp_2)$ consists of ordered pairs of points of $\mathcal{X}(1)$ related by a "$\wp_2$-isogeny"; choosing one of these points yields the degree-3 map $\pi_0 : \mathcal{X}_0(\wp_2) \to \mathcal{X}(1)$. We next determine the ramification of this map. In general, the map $\mathcal{X}_0(I) \to \mathcal{X}(1)$ is branched only above elliptic points of $\mathcal{X}(1)$, if a point P of $\mathcal{X}_0(I)$ above an elliptic point of order e parametrizes an isogeny to some other point of order e' then the ramification index at P is the denominator of the fraction e'/e. [A non-elliptic point is taken to have order 1.] We may regard $\mathcal{X}_0(\wp_2)$ as a symmetric $(3, 3)$ correspondence on $\mathcal{X}(1) \times \mathcal{X}(1)$. We then see that the point $J = \infty$ of order 12 must correspond to the point $J = 0$ of order 4 with multiplicity 3; the point $J = 0$ corresponds to $J = 1$ doubly and $J = \infty$ singly; and $J = 1$ corresponds to $J = 0$ singly and some other point doubly. No other points of $\mathcal{X}(1)$ are ramified in $\mathcal{X}_0(\wp_2)$. Thus by the Riemann-Hurwitz formula $\mathcal{X}_0(\wp_2)$ is again a rational curve, and J is a function of degree 3 with a triple pole such that J and $J-1$ both have double zeros. Up to $\mathrm{Aut}(\mathbf{P}^1)$ there is a unique such function; we choose a rational coordinate t on $\mathcal{X}_0(\wp_2)$ such that $J = t(4t-3)^2$ (so $J-1 = (t-1)(4t-1)^2$). Then the involution[6] $w^{(1)}$ must interchange the points $t=0$, $t=\infty$ parametrizing isogenies between $J=0$ and $J=\infty$, and the points $t=1$, $t=3/4$ parametrizing isogenies between $J=0$ and $J=1$. Therefore $w^{(1)}(t) = 3/4t$.

---

[5] Note that this is not possible with elliptic modular curves, or for that matter with Drinfeld modular curves, precisely because of their cusps. However, the ramification in towers of elliptic or Drinfeld modular curves is small enough to be captured by a tower of ray class fields, suggesting that ray class-field towers might be a fruitful source of curves with many points even over finite fields of non-square order.

Now the curve $\mathcal{X}_0(\wp_2^2)$ covers $\mathcal{X}_0(\wp_2)$ with degree 2, and the only branch points are $t=\infty$ and $t=3/4$. Thus $\mathcal{X}_0(\wp_2^2)$ is again a rational curve, and we may choose a rational coordinate ξ for it such that $t = (\xi^2+3)/4$. Of the points $\xi = \pm 1$ above $t=1$, one must parametrize a $\wp_2^2$-isogeny from $J=1$ to $J=\infty$, the other an isogeny from $J=1$ to itself; we choose ξ so the former point is $\xi = 1$. Then $w^{(2)}$ must switch that point with the point $\xi=\infty$, and fix the other point $\xi=-1$; therefore this involution is $\xi \longleftrightarrow (\xi+3)/(\xi-1)$. As a further check on the computation, note that this involution also switches the two points $\xi = \pm\sqrt{-3}$ above $t=0$, parametrizing a $\wp_2^2$-isogeny from $J=0$ to itself.

We now have all the information needed to determine the Shimura modular curve $\mathcal{X}_0(\wp_2^n)$ for all $n>1$: that curve has $n-1$ coordinates $x_1, \ldots, x_{n-1}$, satisfying the $n-2$ relations

$$\left(\frac{x_j^2+3}{4}\right)\left(\frac{z_{j+1}^2+3}{4}\right) = \frac{3}{4}, \tag{46}$$

that is,

$$(x_j^2+3)(z_{j+1}^2+3) = 12 \qquad (j = 1, \ldots, n-2), \tag{47}$$

where

$$z_j := (x_j+3)/(x_j-1), \tag{48}$$

the same involution we used in (17) for the tower of classical modular curves $\mathrm{X}_0(2^n)$. Unlike these curves, though, the Shimura tower $\mathcal{X}_0(\wp_2^n)$ turns out to be unramified past $n=5$, as may be seen either directly from the formulas (47,48) or from the general description of ramification in the map $\mathcal{X}_0(I) \to \mathcal{X}(1)$. Since each step in the tower is a cyclic extension, it follows that over any finite field of odd characteristic the tower is dominated by the 2-class-field tower of the curve $\mathcal{X}_0(\wp_2^5)$.

For our second example, we choose for K the cubic field $\mathbf{Q}(2\cos\pi/9)$ and for A the quaternion algebra ramified only at two of the three infinite places of K. Then we find in [T] that the group of units of norm 1 in A is the $(2,3,9)$ triangle group. We exhibit the tower $\{\mathcal{X}_0(\wp_3^n)\}_{n>1}$, where $\wp_3$ is the prime of K of residue field $\mathbf{F}_3$. The equations were obtained in the same way that we found (47,48); we leave the intermediate steps as an exercise. Again we find formulas similar to those we obtained earlier (18,19) for the classical modular curves: there are $n-1$ coordinates $x_1, \ldots, x_{n-1}$, related by $n-2$ equations

$$x_j^3 + z_{j+1}^3 = 1 \qquad (j = 1, \ldots, n-2) \tag{49}$$

(this time even simpler than the equation (18) for the classical case), where again

$$z_j := (x_j+2)/(x_j-1). \tag{50}$$

Again the tower has cyclic steps and is unramified after finitely many steps; we find that it is dominated by the 3-class-field tower of the curve $\mathcal{X}_0(\wp_3^4)$.

---

[6] We suppress the unwieldy subscript $\wp_2$.

Fantasia: a speculation on modularity. All our towers are of the following form: the bottom curve C_1 over some finite field k is equipped with an irreducible correspondence $\Phi \subset C_1 \times C_1$ of bidegree (l, l) and a set $S \subset C_1(k)$ of rational points each of which corresponds under Φ with l *distinct* points also in S; the n-th curve in the tower is then the curve C_n of n-tuples $(P_1, \ldots, P_n) \in C_1^n$ such that $(P_j, P_{j+1}) \in \Phi$ for $j = 1, 2, \ldots, n-1$. Then C_n has at least $l^{n-1}|S|$ rational points, and at least when Φ is tamely ramified we can find the genus of C_n as a function of n. For instance, if $C_1 = \mathrm{X}_0(l^2)$, Φ is the image of $\mathrm{X}_0(l^3)$ under $\pi_0 \times \pi_1$, and S is the set of supersingular points, then we recover the tower of curves $\mathrm{X}_0(l^{n+1})$ of our Proposition.

But the (C_1, Φ, S) description makes no assumption of modularity: we can, as in [GS2], try any C_1 and Φ and hope to find an S that yields many points on C_n. In fact, several such (C_1, Φ) were found to admit S large enough to make the tower $\{C_n\}$ asymptotically optimal [GS1,GS2]. However, in each such case $\{C_n\}$ was subsequently explained as a modular tower.

This leads us to speculate: perhaps every asymptotically optimal tower of this recursive form must be modular?

Acknowledgements. Thanks to H. Stichtenoth and P. Solé for e-mail correspondence on the explicit towers and AGM connections eventually published in [GS2] and [S1] respectively. Thanks also to J. Cazaran for reading the manuscript and spotting several local errors as the paper was going into print. Symbolic computations were greatly facilitated by the computer packages PARI and MACSYMA.

This work was made possible in part by funding from the National Science Foundation and the Packard Foundation.

References

[E] Elkies, N.D.: Elliptic and modular curves over finite fields and related computational issues. To appear in the proceedings of the conference *Computational Perspectives on Number Theory* held September, 1995 in Chicago in honor of A.O.L. Atkin, J. Teitelbaum, ed.

[G] Goppa, V.D.: Codes on algebraic curves, *Soviet Math. Dokl.* **24** (1981) #1, 170–172.

[GS1] Garcia, A., Stichtenoth, H.: A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound. *Invent. Math.* **121** #1 (1995), 211–233.

[GS2] Garcia, A., Stichtenoth, H.: Asymptotically good towers of function fields over finite fields. *C. R. Acad. Sci. Paris I* **322** (1996), 1067–1070.

[I] Ihara, Y.: Some remarks on the number of rational points of algebraic curves over finite fields. *J. Fac. Sci. Tokyo* **28** (1981), 721–724.

[S1] Solé, P.: D_4, E_6, E_8 and the AGM. Pages 448–455 in *Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes: 11th International Symposium (AAECC-11), Paris, France, July 1995* (G. Cohen, M. Giusti, T. Mora, eds.; Springer LNCS **948**).

[S2] Solé, P.: Towers of Function Fields and Iterated Means. Preprint, 1997.

[T] Takeuchi, K.: Commensurability classes of arithmetic triangle groups. *J. Fac. Sci. Univ. Tokyo* **24** (1977), 201–212.

[TVZ] Tsfasman, M.A., Vlăduţ, S.G., Zink, T.: Modular curves, Shimura curves and Goppa codes better than the Varshamov-Gilbert bound. *Math. Nachr.* **109** (1982), 21–28.

On the missing functions of a pyramid of curves

Ruud Pellikaan

Department of Mathematics and Computing Science,
Technical University of Eindhoven,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

Abstract

We consider an asymptotically good tower $(\mathcal{Y}_m)_{m \geq 1}$ of curves over a finite field $\mathbb{F}_q$, where q is a square. The Weierstrass semigroup H_m of a rational point $P_\infty^{(m)}$ of $\mathcal{Y}_m$ is determined. For $n \in H_m$ and $m = 1, 2$ and 3 we will give rational functions on $\mathcal{Y}_m$ that have pole order n at $P_\infty^{(m)}$ and that have no poles outside $P_\infty^{(m)}$. For larger m an explicit description of all these functions is still missing.

1 Introduction

A tower

$$\cdots \to \mathcal{X}_m \to \mathcal{X}_{m-1} \to \cdots \to \mathcal{X}_3 \to \mathcal{X}_2 \to \mathcal{X}_1$$

of algebraic curves over a finite field $\mathbb{F}_q$ is said to be *asymptotically good* if

$$\lim_{m \to \infty} \frac{\text{number of rational points of } \mathcal{X}_m/\mathbb{F}_q}{\text{genus of } \mathcal{X}_m} > 0.$$

Recently an explicit description was obtained of several asymptotically good towers by Garcia and Stichtenoth The motivation to consider these came from coding theory: such towers give rise to asymptotically good sequences of codes if this limit is larger than one. Although the existence of good codes on or above the Tsfasman-Vlăduţ-Zink bound was guaranteed and even a polynomial construction was given, the methods used (namely, modular curves) and the degree of the complexity of the construction were such that hardly any of the resulting codes were known explicitly. Now that asymptotically good towers $(\mathcal{X}_m | m \geq 1)$ of algebraic curves are known explicitly, the next step would be to give an explicit description of the vector spaces $\mathcal{L}(G_m)$ resp. $\mathcal{L}(nP_\infty^{(m)})$, where G_m is a divisor (resp. $P_\infty^{(m)}$ is a rational point of $\mathcal{X}_m$. The latter space $\mathcal{L}(nP_\infty^{(m)})$ is the $\mathbb{F}_q$-vector space of all rational functions on $\mathcal{X}_m$ that have no poles outside $P_\infty^{(m)}$ and pole order at most n at $P_\infty^{(m)}$. The first attempts have been made in this direction.

2 Weierstrass semigoups in a tower of curves

In this paper we consider the tower $(\mathcal{Y}_m)_{m\geq 1}$ over $\mathbb{F}_q$, where $q = r^2$ and r is a prime power. It is defined by giving the corresponding function fields $(T_m | m \geq 1)$ as follows:

$$T_1 = \mathbb{F}_q(x_1) \text{ and } T_{i+1} = T_i(x_{i+1}) \text{ with } x_{i+1}^r + x_{i+1} = \frac{x_i^r}{x_i^{r-1}+1}.$$

By the form of the defining equations it is readily seen that $N(\mathcal{Y}_m)$, the number of rational points of $\mathcal{Y}_m$, is at least $(r^2 - r)r^{m-1}$. The genus $g(\mathcal{Y}_m)$ is computed by the theory of Artin-Schreier extensions, and one finds that

$$\lim_{m\to\infty} \frac{N(\mathcal{Y}_m)}{g(\mathcal{Y}_m)} = r - 1 = \sqrt{q} - 1.$$

Hence the tower is asymptotically good and in fact optimal; it attains the Drinfeld-Vlăduţ bound. This implies that geometric Goppa codes which are constructed by means of this tower lie on or above the TVZ bound, which is better than the Gilbert-Varshamov bound for all $q = r^2 \geq 49$.

The element $x_1 \in T_1 \subseteq T_m$ has in T_m a unique pole at the rational point that we denote by $P_\infty^{(m)}$ or by P_∞ for short. Hence it is natural to consider the spaces $\mathcal{L}(nP_\infty^{(m)})$ for all m and n. The dimension of these spaces is determined. The *Weierstrass semigroup* H_m of $P_\infty^{(m)}$, that is $H_m = H(P_\infty^{(m)})$, is the set of all $n \in \mathbb{N}_0$ such that there is some $f \in T_m$ having a pole of order n at $P_\infty^{(m)}$ and no pole outside $P_\infty^{(m)}$.

Theorem 2.1 *Let*

$$c_m = \begin{cases} r^m - r^{\frac{m}{2}} & \text{if } m \text{ is even,} \\ r^m - r^{\frac{m+1}{2}} & \text{if } m \text{ is odd.} \end{cases}$$

Then $H_1 = \mathbb{N}_0$ *and*

$$H_m = rH_{m-1} \cup \{\, n \in \mathbb{N}_0 \mid n \geq c_m \,\},$$

for all $m \in \mathbb{N}_0$, $m > 0$.

The number c_m is called the *conductor* of H_m.

Let $T_{i,j}$ be the function field of the curve $\mathcal{Y}_{i,j}$ defined by

$$T_{i,j} = \mathbb{F}_q(x_i, x_{i+1}, \ldots, x_{i+j-1}) \subset T_{i+j-1}$$

for $i, j \in \mathbb{N}$. Then $\mathcal{Y}_{1,m} = \mathcal{Y}_m$ and $\mathcal{Y}_{i,m}$ is isomorphic to $\mathcal{Y}_m$ for all i. The inclusions $T_{i,j-1} \subset T_{i,j}$ and $T_{i+1,j-1} \subset T_{i,j}$ give corresponding maps of algebraic curves $\mathcal{Y}_{i,j} \to \mathcal{Y}_{i,j-1}$ and $\mathcal{Y}_{i,j} \to \mathcal{Y}_{i+1,j-1}$, respectively.

The proof is by induction on m and by considering the whole pyramid of curves $\mathcal{Y}_{i,j}$. See [1].

3 Missing functions in a pyramid of curves

The function

$$\prod_{i=1}^{j} (x_i^{r-1} + 1)x_{j+1}^{e}$$

in T_m has a pole of order $r^m - r^{m-j} + er^{m-j-1}$ at $P_\infty^{(m)}$ and has no poles outside $P_\infty^{(m)}$, for all $0 \leq e < r$. But not all values of H_m are obtained in this way.

For $n \in H_m$ and $m = 1, 2$ and 3 we will give functions in T_m that have pole order n at $P_\infty^{(m)}$ and that have no poles outside $P_\infty^{(m)}$. For larger m an explicit description of all these functions is still missing.

Example 3.1 For $\mathcal{Y}_1$ is the projective line and $P_\infty^{(1)} = P_\infty$ is the point at infinity. So $T_1 = \mathbb{F}_q(x_1)$, $H_1 = \mathbb{N}_0$. The function x_1^n has a pole of order n at P_∞ and has no poles outside P_∞.

Example 3.2 Let $x = x_1$ and $y = x_2$. Then $T_2 = \mathbb{F}_q(x, y)$ where

$$y^r + y = \frac{x^r}{x^{r-1}+1}.$$

The pole $P_\infty^{(1)}$ of x on $\mathcal{Y}_1$ is totally ramified in the cover $\mathcal{Y}_2$ of $\mathcal{Y}_1$, and $P_\infty^{(2)}$ lies above $P_\infty^{(1)}$. So x has a pole of order r at $P_\infty^{(2)} = P_\infty$.

Define $\Omega = \{\alpha \in \mathbb{F}_q \mid \alpha^r + \alpha = 0\}$ and $\Omega^* = \Omega \setminus \{0\}$. Then Ω and Ω^* consist of r and $r-1$ elements, respectively, since $q = r^2$. For every $\alpha \in \Omega^*$ the function $x - \alpha$ has exactly one zero on $\mathcal{Y}_2$ that we denote by P_α. For any $\beta \in \Omega$, there is a unique common zero of x and $y - \beta$ that we denote by Q_β. The points P_∞ and P_α are poles of $y - \beta$ for all $\alpha \in \Omega^*$ and $\beta \in \Omega$. Hence we have the following principal divisors

$$(x) = \Sigma_{\beta\in\Omega} Q_\beta - rP_\infty,$$

$$(x - \alpha) = rP_\alpha - rP_\infty \qquad \text{if } \alpha \in \Omega^*,$$

$$(y - \beta) = rQ_\beta - \Sigma_{\alpha\in\Omega^*} P_\alpha - P_\infty \qquad \text{if } \beta \in \Omega.$$

Therefore x^i has a pole of order ri at P_∞ and no poles outside P_∞ for all $i \in \mathbb{N}_0$. Furthermore

$$((x^{r-1}+1)y^j) = rjQ_0 + (r-j)\Sigma_{\alpha\in\Omega^*} P_\alpha - [r(r-1)+j]P_\infty.$$

So $(x^{r-1}+1)x^iy^j$ has a pole of order $r(r-1)+ri+j$ at P_∞ and no poles outside P_∞ if $j \leq r$. In this way the functions x^i, $i \in \mathbb{N}_0$ and $(x^{r-1}+1)x^iy^j$, $i, j \in \mathbb{N}_0$, $0 < j < r$ give all possible pole orders of the Weierstrass semigroup

$$H_2 = \{0, r, 2r, \ldots, (r-1)r, (r-1)r+1, (r-1)r+2, \ldots\}.$$

Finally we will treat T_3. Let $z = x_3$ in T_3. Then $T_3 = T_2(z)$ where

$$z^r + z = \frac{y^r}{y^{r-1}+1}.$$

So we have the following tower and pyramid of function fields

$$\begin{tikzcd}
T_3 & & & \mathbb{F}_q(x,y,z) & & \\
T_2 \arrow[u] & & \mathbb{F}_q(x,y) \arrow[ur] & & \mathbb{F}_q(y,z) \arrow[ul] & \\
T_1 \arrow[u] & \mathbb{F}_q(x) \arrow[ur] & & \mathbb{F}_q(y) \arrow[ul] \arrow[ur] & & \mathbb{F}_q(z). \arrow[ul]
\end{tikzcd}$$

There is a unique point P of $\mathcal{Y}_3$ that is totally ramified above P_∞ or P_α, $\alpha \in \Omega^*$, since these points are simple poles of y on $\mathcal{Y}_2$. Denote this point again by P_∞ or P_α, respectively.

Therefore we have the following principal divisor on $\mathcal{Y}_3$

$$(x-\alpha) = r^2 P_\alpha - r^2 P_\infty \quad \text{if } \alpha \in \Omega^*.$$

Hence

$$(x^{r-1}+1) = r^2 \Sigma_{\alpha\in\Omega^*} P_\alpha - r^2(r-1)P_\infty \quad \text{if } \alpha \in \Omega^*.$$

Let $\beta \in \Omega$. Define $\Omega(\beta) = \{\gamma \mid \gamma^r + \gamma = \beta^r\}$.

Proposition 3.3 *Let $\beta \in \Omega$. Then there exists exactly one point $Q_{\beta\gamma}$ on $\mathcal{Y}_3$ above Q_β such that*

$$z - \frac{\beta^{r+1}}{x} - \gamma$$

has a zero in $Q_{\beta\gamma}$ and $v_{Q_{\beta\gamma}}(y-\beta) = r$ for all $\gamma \in \Omega(\beta)$.

We refer for the proof to [2].

A consequence of this proposition is that $\{Q_{\beta\gamma} \mid \gamma \in \Omega(\beta)\}$ is the set of points of $\mathcal{Y}_3$ lying above Q_β. So Q_β is unramified and

$$(x) = \Sigma_{\beta\in\Omega,\ \gamma\in\Omega(\beta)} Q_{\beta\gamma} - r^2 P_\infty,$$

$$(y-\beta) = r\Sigma_{\gamma\in\Omega(\beta)} Q_{\beta\gamma} - r\Sigma_{\alpha\in\Omega^*} P_\alpha - rP_\infty \quad \text{if } \beta \in \Omega,$$

$$(z-\gamma) = r^2 Q_{0\gamma} - \Sigma_{\beta\in\Omega^*,\ \gamma'\in\Omega(\beta)} Q_{\beta\gamma'} - \Sigma_{\alpha\in\Omega^*} P_\alpha - P_\infty \quad \text{if } \gamma \in \Omega.$$

Definition 3.4 Let t be a local parameter at a point Q. Then we write

$$f = g + \mathcal{O}(t^e) \ \text{ if } v_Q(f-g) \geq e.$$

The function x is a local parameter at $Q_{\beta\gamma}$ and

$$y = \beta + \mathcal{O}(x^r) \text{ and } z = \frac{\beta^{r+1}}{x} + \mathcal{O}(1),$$

by Proposition 3.3. Furthermore $\beta^{r+1} = -\beta^2$ for all $\beta \in \Omega$. Hence

$$z + \frac{y^2}{x} = \mathcal{O}(1)$$

at $Q_{\beta\gamma}$ and for all $\beta \in \Omega$ and $\gamma \in \Omega(\beta)$. It still has poles outside P_∞ but these are cancelled by the zeros of $x^{r-1} + 1$. Therefore

$$(x^{r-1} + 1)\left(z + \frac{y^2}{x}\right)$$

is a function that has no poles outside P_∞. The functions

$$(x^{r-1} + 1)y^j\left(z + \frac{y^2}{x}\right)^k$$

have no poles outside P_∞ if $j + 2k \leq r$. Now the trick is to expand $\beta^j(z + \frac{\beta^2}{x})^k$ first by Newton's binomium, see which terms β^{2i+j} appear, replace this by a term with the right coefficient $c(i,j)$ and the right exponent $e(i,j)$ modulo $r-1$, and use these coefficients and exponents with y instead of β.

Definition 3.5 Let $e(0,j) = j$. Let $e(i,j)$ be the integer in the interval $[1, r-1]$ that is congruent to $2i + j$ modulo $r - 1$, if $i > 0$.

Definition 3.6 Define the constant $c(i,j)$ by

$$c(i,j) = (-1)^{\lfloor \frac{2i+j-1}{r-1} \rfloor}.$$

Remark 3.7 Notice that

$$\beta^{2i+j} = c(i,j)\beta^{e(i,j)},$$

since $\beta^r = -\beta$ for all $\beta \in \Omega$ and by the definitions of $e(i,j)$ and $c(i,j)$. Furthermore

$$c(i,j)y^{e(i,j)} = \beta^{2i+j} + \mathcal{O}(x^r)$$

at $Q_{\beta\gamma}$, since $e(i,j) < r$ and $y = \beta + \mathcal{O}(x^r)$.

Definition 3.8 Let j, k be integers such that $0 \leq j, k < r$. Define the function $f_{jk} \in T_3$ by

$$f_{jk} = \sum_{i=0}^{k} \binom{k}{i} c(i,j) \frac{y^{e(i,j)}}{x^i} z^{k-i}$$

Lemma 3.9 *The function f_{jk} has no pole at $Q_{\beta\gamma}$ for all $\beta \in \Omega$, $\gamma \in \Omega(\beta)$ and $j, k < r$.*

Proof. We will show that $x^k f_{jk} = \mathcal{O}(x^k)$ at $Q_{\beta\gamma}$.

$$\begin{aligned} x^k f_{jk} &= \textstyle\sum_{i=0}^{k} \binom{k}{i} c(i,j) y^{e(i,j)} (xz)^{k-i} \\ &= \textstyle\sum_{i=0}^{k} \binom{k}{i} \beta^{2i+j} (xz)^{k-i} + \mathcal{O}(x^r) \\ &= y^j \left(\textstyle\sum_{i=0}^{k} \binom{k}{i} \beta^{2i} (xz)^{k-i} \right) + \mathcal{O}(x^r) \\ &= y^j (xz + \beta^2)^k + \mathcal{O}(x^r) = x^k y^j (z + \tfrac{\beta^2}{x})^k + \mathcal{O}(x^r) = \mathcal{O}(x^k) \end{aligned}$$

The second equality is Remark 3.7. Proposition 3.3 is used in the third and sixth equality for y and z, respectively. The fourth equality is Newton's binomium. □

Proposition 3.10 *The rational function $(x^{r-1} + 1) f_{jk}$ on $\mathcal{Y}_3$ has a pole at P_∞ of order $r^2(r-1) + jr + k$ and no poles outside P_∞ for all $j, k < r$.*

Proof. To investigate the poles of these functions it is enough to consider the points P_∞, P_α, $\alpha \in \Omega^*$ and the points $Q_{\beta\gamma}$. In Lemma 3.9 it is shown that f_{jk} has no pole at the $Q_{\beta\gamma}$, and the same holds for $x^{r-1} + 1$. So it remains to investigate the behaviour at P_α and P_∞.

If $i \leq k < r$ and $j < r$, then

$$v_{P_\alpha}\left(\frac{y^{e(i,j)}}{x^i} z^{k-i} \right) = -e(i,j)r - (k-i) - i.0 \geq -(r-1)r - (r-1) = -r^2 + 1,$$

since $e(i,j) \leq r - 1$. Hence the pole order of f_{jk} is at most $r^2 - 1$ at P_α. But $(x^{r-1} + 1)$ has a zero of order $r^2 - 1$ at P_α. So $(x^{r-1} + 1) f_{jk}$ has no pole at P_α.

Now the pole order at P_∞ is considered. If $i \geq 1$, then

$$v_{P_\infty}\left(\frac{y^{e(i,j)}}{x^i} z^{k-i} \right) = -e(i,j)r - (k-i) - i.(-r^2) \geq -(r-1)r - (r-1) + r^2 \geq 1,$$

since $e(i,j) \leq r-1$. If $i = 0$, then

$$v_{P_\infty}\left(\frac{y^{e(i,j)}}{x^i} z^{k-i}\right) = v_{P_\infty}(y^j z^k) = -jr - k.$$

The pole order of $(x^{r-1}+1)$ is $r^2(r-1)$ at P_∞. Therefore $(x^{r-1}+1)f_{jk}$ has a pole of order $r^2(r-1) + jr + k$ at P_∞. □

Corollary 3.11 *Let R_3 be the subring of T_3 consisting of all rational functions that have no poles outside P_∞. Then*

$$\{x^i \mid \in \mathbb{N}_0,\ i < r-1\} \cup \{\ (x^{r-1}+1)x^i f_{jk} \mid i,j,k \in \mathbb{N}_0,\ j,k < r\ \}$$

is a basis of R_3 over $\mathbb{F}_q$.

Proof. These functions are elements of R_3 and the pole orders of these functions form the Weierstrass semigroup H_3 of P_∞, by Theorem 2.1 and Proposition 3.10, and are mutually distinct. □

Example 3.12 Let $q = 2$. Then the conductor is $c_3 = 2^3 - 2^2 = 4$ and $H_3 = \{0\} \cup \{n \in \mathbb{N}_0 \mid n \geq 4\}$. The functions in R_3 with pole orders $4, 5, 6$ and 7 are $x+1$, $(x+1)(z+y/x)$, $(x+1)y$ and $(x+1)y(z+1/x)$, respectively.

Example 3.13 Let $q = 3$. Then the conductor is $c_3 = 3^3 - 3^2 = 18$ and $H_3 = \{0, 9\} \cup \{n \in \mathbb{N}_0 \mid n \geq 18\}$. The functions in R_3 with pole orders between 18 and 26 are in increasing order x^2+1, $(x^2+1)(z+y^2/x)$, $(x^2+1)(z^2 - y^2/x - y^2/x^2)$, $(x^2+1)y$, $(x^2+1)y(z-1/x)$, $(x^2+1)y(z-1/x)^2$, $(x^2+1)y^2$, $(x^2+1)y^2(z-1/x)$ and $(x^2+1)y^2(z-1/x)^2$, respectively.

References

[1] R. Pellikaan, H. Stichtenoth and F. Torres, "Weierstrass semigroups in an asymptotically good tower of function fields," to appear in *Finite Fields and their Applications.*

[2] A. Garcia and H. Stichtenoth, "On the asymptotic behaviour of some towers of function fields over finite fields," *J. Number Theory*, vol. 61, pp. 248-273, 1996.

Quadratic Residue Codes and Hyperelliptic Curves

Iwan M. Duursma
duursma@research.att.com

For the quadratic residue code of length a prime number $p \equiv 3 \pmod 4$ defined over the alphabet $Z/\ell Z$, we give a geometric interpretation of the group of syndromes. We show that the group can be identified with a ℓ-torsion subgroup of the divisor class group of the curve $y^2 = x^p - x$. We show that ℓ-torsion points on the elliptic curve $y^2 = x^3 - x$ are a natural choice for the alphabet of the code. Our future goal is to derive new properties of quadratic residue codes from the arithmetic properties of the curve $y^2 = x^p - x$. Another goal is to explore the use of points on an elliptic curve as alphabet for codes.

We recall the construction of the quadratic residue code of length p over the field F of size ℓ, where p and ℓ are distinct prime numbers such that ℓ is a square modulo p [1]. Let $F(\zeta)$ be the cyclotomic extension of F generated by a primitive p-th root of unity ζ. We assume that $p \equiv 3 \pmod 4$, which is essential for our geometric construction. Let Q be the set of nonzero quadratic residues modulo p; so that the set $-Q$ consists of the nonzero quadratic non-residues. Let $\bar{C}$ be the code over $F(\zeta)$ that is orthogonal to the vectors

$$(1, \zeta^i, \zeta^{2i}, \ldots, \zeta^{(p-1)i}), \qquad i \in Q.$$

Since $\ell \in Q$, the code $\bar{C}$ has a set of generators defined over the field F. We define the quadratic residue code C over F as the code generated by these generators, or $C \overset{\text{def}}{=} \bar{C} \cap (F \times F \times \cdots \times F)$. For $p \equiv 3 \pmod 4$, the code C contains its dual code, $C = C^\perp + (1, 1, \ldots, 1)$. We assume furthermore that $4\ell | p+1$, which leads to a simpler presentation. This includes as special cases the Hamming code $[7,4,3]$ and the Golay codes of type $[11,6,5]$ and $[23,12,7]$, respectively, for $(p = 7, \ell = 2)$, $(p = 11, \ell = 3)$ and $(p = 23, \ell = 2)$, respectively.

The following properties can be found in [1]. For a codeword $(c_0, c_1, \ldots, c_{p-1}) \in C$ and for $i \in Q$,

$$c_o + c_1\zeta^i + c_2\zeta^{2i} + \ldots + c_{p-1}\zeta^{(p-1)i} = 0.$$

The codewords of C can therefore be described as the coordinate vectors of elements in the ideal

$$I \overset{\text{def}}{=} (\prod_{i \in Q}(X - \zeta^i)) \subset F[X]/(X^p - 1)$$

on the basis $(1, X, X^2, \ldots, X^{p-1})$. The ideal I has a unique idempotent e; that is, $I = (e)$ and $e^2 = e$. There exist $a, b, c \in F$ such that

$$e = a + b\sum_{i \in Q} X^i + c\sum_{i \in -Q} X^i.$$

For $p \equiv 3 \pmod 4$,

$$(x - \sum_{i \in Q}\zeta^i)(x - \sum_{i \in -Q}\zeta^i) = x^2 + x + (p+1)/4.$$

It follows that, for $4\ell|(p+1)$,

$$(e, e^\perp) = (\sum_{i \in Q} X^i, 1 + \sum_{i \in -Q} X^i),\ (\sum_{i \in -Q}, 1 + \sum_{i \in Q} X^i).$$

The two choices for the pair of idempotents yield two different codes that differ by a reversal of the codewords. With minor modifications, the results apply to quadratic residue codes over a ring $Z/\ell Z$ [2]. For a geometric and different construction of quadratic residue codes, let E be the elliptic curve $y^2 = x^3 - x$ over the field F_p of p elements. For $p \equiv 3 \pmod 4$, the curve is supersingular with $p+1$ rational points and with group structure,

$$E(F_p) = Z/2\ell Z \times Z/2Z, \qquad \ell = (p+1)/4.$$

The group has canonical subgroups $2E = Z/\ell Z$ and $E[2] = Z/2Z \times Z/2Z$. The group $2E$ gives us an alphabet $Z/\ell Z$. Codes over a smaller alphabet follow immediately by reduction. Codes over a larger alphabet require an extension of the base field F_p of E. Some explicit cases:

$$\begin{array}{ll} p = 7: & 2E = \{O, (1,0)\}. \\ p = 11: & 2E = \{O, (4,4), (4,7)\}. \\ p = 23: & 2E = \{O, (3,1), (2,12), (1,0), (2,11), (3,22)\}. \end{array}$$

As before, we need an extension of the alphabet to define a cyclic code of length p. We use a covering of the curve E by the curve $K : y^2 = x^p - x$. Let τ be an automorphism $(x,y) \longmapsto (\zeta^2 x, -\zeta y)$ of K of order $(p-1)/2$. It has invariants $v = yx^{(p-3)/4}$ and $u = x^{(p-1)/2}$, such that

$$v^2 = y^2 x^{(p-3)/2} = x^{(3p-3)/2} - x^{(p-1)/2} = u^3 - u.$$

Thus, the extension K/E is cyclic of degree $(p-1)/2$. The important point is that K also has an automorphism $\sigma : (x,y) \mapsto (x+1, y)$ of order p; just as $F_\ell(\zeta)$ before had an automorphism of order p.

$$\begin{tikzcd} F_\ell(\zeta) \arrow[d, dash] & \sigma : x \mapsto \zeta x. & K/F_p : \; y^2 = x^p - x \arrow[d, dash] & \sigma : x \mapsto x+1. \\ F_\ell & & E/F_p : \; v^2 = u^3 - u & \end{tikzcd}$$

Ramification occurs above the points P_∞ and $(0,0)$ of E. The ramification is complete in both cases. The other 2-torsion points $(1,0)$ and $(-1,0)$ split completely, giving a total of $p+1$ rational points on K. The remaining rational points of E are inert. Above a finite point (a,b) of E with $a \neq 0$, lie the $(p-1)/2$ points $(x, bx^{(p+1)/4}/a)$, for $x^{(p-1)/2} = a$. The divisor class group Γ of K is of type

$$\Gamma(F_p) = (Z/2\ell Z \times Z/2Z)^{(p-1)/2}.$$

There is a natural isomorphism from the invariant subgroup Γ/τ of Γ to E, which maps

$$\begin{array}{lcl} [1,0] & \longmapsto & [1,0], \\ [x^{(p-1)/2}, 0] & \longmapsto & [u,0], \\ [x^{(p-1)/2} - a, bx^{(p+1)/4}/a] & \longmapsto & [u-a, b]. \end{array}$$

The notation is taken from Cantor's algorithm for addition in the divisor class group of a hyperelliptic curve [3]. We can now describe the code construction. Choose a point $(a,b) \in 2E$ of order ℓ and let $P_0 \in 2\Gamma$ be the element that corresponds to it under the above bijection. The action of σ gives p distinct group elements $P_i = \sigma^i(P_0) \in 2\Gamma$, for $i = 0, 1, \ldots, p-1$.

Theorem: The quadratic residue code C of length $p = 4\ell - 1$ over the alphabet $Z/\ell Z$ can be defined as the set of all codewords $(c_0, c_1, \ldots, c_{p-1}) \in (Z/\ell Z)^p$, such that

$$c_0 P_0 + c_1 P_1 + c_2 P_2 + \ldots + c_{p-1} P_{p-1} = 0.$$

Proof: It follows from the action of σ that the code C is cyclic of length p. The automorphism τ acts on the elements $P_0, P_1, \ldots, P_{p-1}$,

$$\begin{aligned} \tau P_i &= ((x/\zeta^2 + i)^{(p-1)/2} - a, -\zeta b(x/\zeta^2 + i)^{(p+1)/4}/a) \\ &= ((x + i\zeta^2)^{(p-1)/2} - a, \zeta^{(p+1)/2} b(x/\zeta^2 + i)^{(p+1)/4}/a) \\ &= ((x + i\zeta^2)^{(p-1)/2} - a, b(x + i\zeta^2)^{(p+1)/4}/a). \end{aligned}$$

Under the action, points P_i divide into three orbits, according to $i = 0, i \in Q, i \in -Q$. The generating polynomial of the code C is built from the factors $(X - 1)$, $\prod_{i \in Q}(X - \zeta^i)$ or $\prod_{i \in -Q}(X - \zeta^i)$. The rank of the group generated by $P_0, P_1, \ldots, P_{p-1} \in 2\Gamma$ is at most the rank $(p-1)/2$ of 2Γ. The rank of the code is therefore at least $(p+1)/2$. Since the code is non-trivial, we have equality for the ranks in both cases. And the code is one of two possible choices for the quadratic residue code. QED

To obtain the ternary Golay code, we choose the point $(4, 4)$ of order three on E, which corresponds the element $P_0(x^5 - 4, x^3) \in 2\Gamma = (Z/3Z)^5$. The element is fixed under τ but the action of σ produces the 11 elements $P_i((x + i)^5 - 4, (x + i)^3) \in 2\Gamma$, for $i = 0, 1, \ldots, 10$. With the theorem, the linear relations

$$c_0 P_0 + c_1 P_1 + c_2 P_2 + \ldots + c_{10} P_{10} = 0$$

among the elements $P_0, P_1, \ldots, P_{10}$ define the ternary Golay code $[11, 6, 5]$.

We are investigating (1) new properties of quadratic residue codes that follow from the description in the theorem, (2) new applications for codes whose alphabet consists of points on an elliptic curve, and (3) new codes whose construction generalizes the one given in this paper.

References

[1] F.J. MacWilliams and N.J.A. Slone, *The Theory of Error-Correcting Codes.* North-Holland, Amsterdam, 1977.

[2] A.R. Calderbank, N.J.A Sloane, Modular and p-adic cyclic codes, *Designs, Codes and Cryptography*, vol.6, no.1, pp.21–35, 1995.

[3] D.G. Cantor, Computing in the Jacobian of a hyperelliptic curve, *Math. Comp.*, vol.48, no.177, pp.95–101, 1987.

On the determination of error values for codes from a class of maximal curves

Ralf Koetter, Coordinated Science Laboratory
University of Illinois
1308 West Main Street
Urbana, Illinois, 61801-2307
email: ralf@csl.uiuc.edu

I Introduction

The decoding of algebraic-geometrical (AG) codes is based on the principle of algebraic decoding, pioneered for the decoding of Reed-Solomon (RS) codes. The decoding algorithm consists of two major steps, namely first finding the error positions as the zeros of a so-called error-locator, and second using this error-locator to find the error values. The first step of finding the error-locator has been subject of active research and efficient methods based on Berlekamp-Massey type algorithms (BMA) to accomplish this task have been developed in e.g. [1][2]. The second step of finding the error values has only recently attracted attention [3][4][5][6]. This may be caused by the fact that the asymptotic complexity of this problem is small in comparison to the first step of the algorithm. Nevertheless as the algorithmic treatment of AG codes matures specific problems such as speed and space complexity become important which the earlier solutions did not address. Here we present a method for calculating error values that generalizes and to a certain extent simplifies Forney's formula for RS codes.

II Problem formulation

Let $\mathbb{F}_q$ denote the finite field with q elements and let

$$\mathcal{X} : F(x,y) = 0$$

be an irreducible, plane curve over $\mathbb{F}_q$ with genus g.

We assume that the affine, $\mathbb{F}_q$-rational points $P_i = (x_i, y_i)$, $i = 1, 2, \ldots, n$ are nonsingular. Let ν_i be the valuation of a rational function on $\mathcal{X}$ at P_i and let P_∞ be a point on $\mathcal{X}$ at infinity. The space of functions with poles only at P_∞ of multiplicity not more than m is denoted by $L(mP_\infty)$. An AG code $C(m)$ may be defined as

$$C(m) = \{c \in \mathbb{F}_q^n : \sum_{i=1}^{n} c_i f(x_i, y_i) = 0, f \in L(mP_\infty)\}.$$

The parameters of $C(m)$ are estimated as length n, dimension $k = n - m + g - 1$ and minimum Hamming distance $d \geq m - 2g + 2$ for $2g - 2 < m < n$ [7]. It is clear that in order to find good codes we need curves with many rational points compared to the genus. The Hasse-Weil bound states that the number of $\mathbb{F}_q$-rational points on $\mathcal{X}$ can not exceed $q + 1 + 2g\sqrt{(q)}$. Curves attaining this bound are called maximal. In this paper we restrict our attention to maximal curves over $\mathbb{F}_{p^{r^2}}$, p prime, of the form $F(x, y) = H(y) - x^{r+1}$ where $H(y)$ is a linearized polynomial

$$H(y) = \sum_{i=0}^{s \leq r} h_i y^{p^i}, \ h_0 = 1.$$

The following theorem is a consequence of a remark in [8].

Theorem 1 *Let $H(y)$ and $M(y)$ be linearized polynomials s.th.*

$$H(M(y)) = y^{p^r} + y.$$

The curve $\mathcal{X} : H(y) = x^{p^r+1}$ is a maximal curve of genus $g = \frac{(p^s-1)p^r}{2}$ where p^s is the degree of $H(y)$.

Proof. The genus of a curve of the above type is well known [9] so we only have to show that the number of $\mathbb{F}_{p^{2r}}$ rational points attains the Hasse-Weil bound. This is well known for the hermitian curve $\mathcal{H} : y^{p^r} + y = x^{p^r+1}$. Let (x_i, y_i) be a point on $\mathcal{H}$. Then $(x_i, M(y_i))$ is a point on $\mathcal{X}$. Let B be the r-dimensional, $\mathbb{F}_p$-linear space of solutions to the equation $y^{p^r} + y$. The polynomial $M(y)$ maps B linearly on a space of dimension $r - dim(kernel(M)) = s$. Thus the p^{3r} affine points on $\mathcal{H}$ are mapped to $p^{2r}p^s$ points on $\mathcal{X}$, which together with one point at infinity satisfies the Hasse-Weil bound with equality. □

Corollary 1 *Let $h(y) = \sum_{i=0}^{s} h_i y^i$ be a factor of $y^r + 1 \pmod{p}$ and let*

$$H(y) = \sum_{i=0}^{s} h_i y^{p^i}.$$

$H(y) - x^{q+1}$ is maximal over $\mathbb{F}_{p^{2r}}$.

Proof. Let $y^r + 1 = h(y)m(y)$ be a factorization of $y^r + 1$. The corresponding linearized polynomials $H(y)$ and $M(y)$ satisfy the conditions of Theorem 1. □

The monomials x and y have pole order p^s and $p^r + 1$ at P_∞ and generate the ring of functions with poles only at P_∞. Thus the codes from these curves may be defined as

$$C(m) := \{\mathbf{c} \in \mathbb{F}_{p^{2r}}^n : \sum_{i=1}^{n} c_i x_i^a y_i^b = 0, ap^s + b(p^r + 1) \leq m, b < p^s\}.$$

Let $\mathbf{r} \in \mathbb{F}_{p^{2r}}^n = \mathbf{c} + \mathbf{e}$, $\mathbf{c} \in C(m)$ be given. After the first step of the decoding algorithm [2] we have at our disposal an error-locating polynomial $\sigma(x, y)$ such that $e_i \sigma(x_i, y_i) = 0$, $i = 1, 2, \ldots, n$. Moreover we are given the syndromes

$$S_{a,b} = \sum_{i=1}^{n} e_i x_i^a y_i^b, \; ap^s + b(p^r + 1) \leq m + g. \tag{1}$$

The problem we are interested in is finding the values e_i.

III Error Values

Let for each point P_i a polynomial $g_i(x, y)$ be defined as

$$g_i(x, y) = \sigma(x, y_i) F(x, y) - \sigma(x, y) F(x, y_i).$$

We quote the following proposition from [4], which states an important property of $g_i(x, y)$. In the proposition $\mathbb{F}_q[x, y]$ denotes the ring of bivariate polynomials with coefficients from $\mathbb{F}_q$ and $f(x, y)$ denotes the curve in consideration.

Proposition 1 *Let ν_Q denote the discrete valuation at Q. Given a polynomial $\sigma(x, y)$, we can construct a function $\sigma' \in \mathbb{F}_q[x, y]$ as*

$$\sigma'(x, y) = \frac{\sigma(x, y) f(x, y_k) - \sigma(x, y_k) f(x, y)}{(y - y_k)(x - x_k)} \tag{2}$$

such that σ' has the property

$$\begin{cases} \nu_{Q_k}(\sigma') = \nu_{Q_k}(\sigma) - 1 \\ \nu_{Q_i}(\sigma') \geq \nu_{Q_i}(\sigma) \quad \text{for all } i \neq k. \end{cases} \tag{3}$$

□

Let the functions $p_i(x,y)$ be defined as

$$p_i(x,y) = \frac{x^{p^{2r}} - x}{\sigma(x,y)} \frac{g_i(x,y)}{(y-y_i)(x-x_i)}.$$

Proposition 2 *The functions $p_i(x,y)$ satisfy the following property*

$$p_i(x_j,y_j) = \begin{cases} 1 & i = j \\ 0 & otherwise. \end{cases} \tag{4}$$

Proof. The function $x^{p^{2r}} - x$ has a simple zero at all affine points P_j, which follows from the fact that $x - x_j$ is never tangent to $\mathcal{X}$. We consider the valuation of $p_i(x,y)$ at points P_i. For $P_j \neq P_i$ we have $\nu_j(p_i(x,y)) = 1 - \nu_j(\sigma(x,y)) + (\geq \nu_j(\sigma(x,y)) \geq 1$ which implies that $p_i(x_j,y_j) = 0$ for $i \neq j$. For $i = j$ we have that $\nu_j(p_i(x,y)) = 1 - \nu_j(\sigma(x,y)) + (\nu_j(\sigma(x,y) - 1) = 0$ and we see that $p_i(x_i,y_i)$ is defined and non-zero. In order to prove that $\nu_i(p_i(x,y))$ equals 1 we note that $p_i(x,y)$ is on $\mathcal{X}$ identical to the function

$$-\frac{(x^{p^{2r}} - x)F(x,y_i)}{(y-y_i)(x-x_i)}$$

From L'Hopitals rule we immediately see that $\frac{(x^{p^{2r}}-x)}{x-x_i}$ equals -1 when evaluated at x_i. the term $\frac{F(x,y_i)}{(y-y_i)}$ may equivalently be written as $H(y-y_i)/(y-y_i)$. Again applying L'Hopitals rule we see that $H(y-y_i)/(y-y_i)$ equals $h_0 = 1$ when evaluated at y_i. □

Let a function $q(x,y)$ now be defined as $q(x,y) = \sum_i e_i p_i(x,y)$. From equation (4) we immediately find the following formula,

$$e_i = q(x_i,y_i).$$

With a function $\omega(x,y)$ defined as

$$\omega(x,y) = \sum_{i=1}^{n} e_i \frac{\sigma(x,y_i)F(x,y) - \sigma(x,y)F(x,y_i)}{(y-y_i)(x-x_i)} \tag{5}$$

we find the following Forney-formula-like expression for the error values e_i

$$e_i = \left. \frac{(x^{p^{2r}} - x)\omega(x,y)}{\sigma(x,y)} \right|_{\substack{x = x_i \\ y = y_i}}.$$

Thus finding the error values comes down to evaluating the above expression. In the following section we turn our attention to the determination of the error-evaluator $\omega(x,y)$.

IV The error evaluator

We will treat the two components of equation (5) separately and denote them by

$$\omega^{(1)}(x,y) = \sum_{i=1}^{n} e_i \frac{\sigma(x,y)F(x,y_i)}{(y-y_i)(x-x_i)} \tag{6}$$

$$\omega^{(2)}(x,y) = \sum_{i=1}^{n} e_i \frac{\sigma(x,y_i)F(x,y)}{(y-y_i)(x-x_i)} \tag{7}$$

Even though $\omega^{(1)}(x,y)$ and $\omega^{(2)}(x,y)$ are in itself not polynomials but rather rational functions it suffices to consider the integral part of $\omega^{(1)}(x,y)$ and $\omega^{(2)}(x,y)$ as the non-integral parts cancel. Let $[\cdot]$ denote the integral part of a rational function. Thus we may write

$$\omega(x,y) = [\omega^{(2)}(x,y)] - [\omega^{(1)}(x,y)].$$

Moreover we write $\sigma(x,y)$ in the form $\sigma(x,y) = \sum_{m=0}^{p^s-1} \sigma_m(x) y^m$. We have the following equations.

$$F(x,y_i)/(x-x_i) = \sum_{a=0}^{p^r} x^a x_i^{q-a}$$

$$[\sigma(x,y)/(y-y_i)] = \sum_{m=1}^{p^s-1} \sigma_i(x)[\frac{y^m}{y-y_i}]$$

$$= \sum_{m=1}^{p^s-1} \sigma_m(x) \sum_{l=0}^{m-1} y^l y_i^{m-1-l} = \sum_{l=0}^{p^s-2} y^l \sum_{m=l+1}^{m-1} \sigma_m(x) y_i^{m-1-l}$$

Substituting these equations into (6) we find

$$[\omega^{(1)}(x,y)] = \sum_{i=1}^{n} e_i \sum_{a=0}^{q} x^a x_i^{q-a} \sum_{l=0}^{p^s-2} y^l \sum_{i=l+1}^{i-1} \sigma_i(x) y_m^{i-1-l},$$

which using equation (1) and some straight forward manipulations leads to

$$[\omega^{(1)}(x,y)] = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} x^a y^b \sum_{l=0}^{\infty} \sum_{s=0}^{\infty} \sigma_{a-l,s+b+1} S_{p^r-l,s}.$$

Numbers with negative indices in this equations are assumed to equal zero. We find the following expression for $[\omega^{(2)}(x,y)]$. The derivation is omitted as it is similar to the derivation of $[\omega^{(1)}(x,y)]$.

$$[\omega^{(2)}(x,y)] = \sum_{k=0}^{s} h_k \sum_{a=0}^{\infty} \sum_{b=0}^{p^k-1} x^a y^b \sum_{l=0}^{\infty} \sum_{j=0}^{\infty} \sigma_{a+l+1,j} S_{l,p^k-1-b+j}$$

V A simplified error-evaluator

The finding of the error locator polynomial in the decoding of AG codes is usually accomplished by a Berlekamp-Massey type algorithm [1][2]. The output of this algorithm is two sets of polynomials $G = \{\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(p^s)}\}$ and $F = \{\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(p^s)}\}$ where F is a Gröbner basis for the error-locating ideal and G is a set of polynomials that failed to satisfy the recursion imposed by the syndrome array at an earlier step of the algorithm. In deriving a simplified error-evaluator we first restrict our attention to RS codes. The approach exhibited there will directly lend itself to a generalization to AG codes. To avoid further notation we simply regard RS codes as AG codes from a curve $y = x^{p^r+1}$. The decoding algorithm degenerates to the standard BMA described e.g. in [10]. Let $\sigma(x)$ be the error-locator found by a BMA for an error vector $\mathbf{e}$ and let $\lambda(x)$ be the polynomial that generated the largest number of initial syndromes $S_a = \sum_{i=1}^{n} e_i x_i^a$. More precisely we have

$$\sum_{j=0}^{\infty} \sigma_j S_{j+a} = 0 \; \forall a \geq 0, \quad \sum_{j=0}^{\infty} \lambda_j S_{j+a} = \begin{cases} 0 & a < \deg(\sigma) - 1 \\ 1 & a = \deg(\sigma) - 1. \end{cases}$$

The following theorem, which is similar to a theorem from [11], shows how to avoid the computation of an error-evaluator in the decoding of RS codes.

Theorem 2 *Let $\sigma(x)$ and $\lambda(x)$ be the output of a BMA for an RS code subject to an error vector* $\mathbf{e}$. *The error values e_i may be found as*

$$e_i = \frac{1}{\lambda(x_i)\sigma_x(x_i)} \text{ if } \sigma(x_i) = 0,$$

where $\sigma_x(x)$ denotes the formal derivative of $\sigma(x)$.

Proof. Consider the modified error vector $\tilde{\mathbf{e}}$ which has elements $e_i \lambda(x_i)$. The vector $\tilde{\mathbf{e}}$ has support in exactly the non-zero positions of $\mathbf{e}$ and gives rise to syndromes $\tilde{S}_a = \sum_{i=1}^{n} e_i \lambda(x_i) x_i^a = \sum_{j=0}^{\infty} \lambda_j S_{j+a}$. The polynomial $\sigma(x)$ is still the error-locator and in order to determine error values $\tilde{e}_i$ we may compute the corresponding error-evaluator $\tilde{\omega}(x)$. The error values $\tilde{e}_i$ are then given as

$$\tilde{e}_i = \frac{\tilde{\omega}(x_i)}{\sigma_x(x_i)} \text{ if } \sigma(x_i) = 0.$$

In Forney's formula the error evaluator $\tilde{\omega}(x)$ may be defined as, see e.g. [2]

$$\tilde{\omega}(x) = \sum_{i=0}^{t-1} x^i \sum_{k=0}^{t-i-1} \sigma_{k+i+1} \tilde{S}_k$$

and we see that $\tilde{\omega}(x)$ equals 1 because all but one of the relevant syndromes $\tilde{S}_k$ are equal to zero . Thus the $\tilde{e}_i$ are given as $1/\sigma_x(x_i)$ and the claim of the theorem follows. □

The situation for AG codes from a curve $\mathcal{X}$ is similar. Let $\sigma(x,y) \in F$ be the error-locator of lowest pole order at P_∞ and let $\lambda(x,y) \in G$ be a polynomial that failed to satisfy the recurrence at an earlier step of the algorithm. More precisely we have as a consequence of the Berlekamp-Massey type algorithm [1][2] for any $\sigma(x,y) \in F$ and $\lambda(x,y) \in G$

$$\sum_{i,j} \sigma_{i,j} S_{i+a,j+b} = 0 \ \forall (a,b)\ 0 \le a,\ 0 \le b \tag{8}$$

$$\sum_{i,j} \lambda_{i,j} S_{i+a,j+b} = \begin{cases} 0 & ap^s + b(p^r+1) < \nu_\infty(\sigma) - p^s - 1 \\ \tilde{S}_{a,b} & ap^s + b(p^r+1) \ge \nu_\infty(\sigma) - p^s - 1. \end{cases} \tag{9}$$

The proof of Theorem 2 suggests to solve for the error values of a modified error vector $\tilde{e} = (e_1\lambda(x_1,y_1), \ldots, e_n\lambda(x_n,y_n))$. In other words we have to find an error-evaluator $\tilde{\omega}(x,y)$ from equations (6) and (7) for syndromes $\tilde{S}_{a,b} = \sum_{i,j}^{\infty} \lambda_{i,j} S_{i+a,j+b}$ rather than for the original syndromes. Given $\lambda(x,y)$ it is easy to find these modified syndromes from (9). Indeed most modified syndromes are 0 and we do not have to determine too many non-zero modified syndromes. For simplicity we assume in the sequel that the number of errors t satisfies $t \ge 4g+2p^s+1$. However it should be clear from the sequel that this is not an essential condition on the proposed method. The error-evaluator $\tilde{\omega}(x,y)$ is again composed of two components $[\tilde{\omega}^{(1)}(x,y)]$ and $[\tilde{\omega}^{(2)}(x,y)]$ which may be determined by equation (6) and (7), where $\tilde{S}_{a,b}$ replaces $S_{a,b}$.

Lemma 1 *Let an error vector* e *with weight* t *larger than* $4g+2p^s+1$ *be given. Then* $[\tilde{\omega}^{(1)}(x,y)] = 0$ *and* $[\tilde{\omega}^{(2)}(x,y)]$ *has at most* $g+p^s-1$ *coefficients. In order to compute* $[\tilde{\omega}^{(2)}(x,y)]$ *we need to determine at most* $2g+p^s+1$ *syndromes* $\tilde{S}_{a,b}$.

Proof. For the claim that $[\tilde{\omega}^{(1)}(x,y)]$ equals zero it suffices to show that all modified syndromes in the calculation of $[\tilde{\omega}^{(1)}(x,y)]$ equal zero. The syndrome corresponding to the monomial with highest pole order in (6) is $\tilde{S}_{p^r,p^s-2}$. If t errors have occurred then the pole order of the error-locator is at least t and it follows from (9) that all syndromes $\tilde{S}_{a,b}$ with $ap^s + b(p^r+1) < t - p^s - 1$ are zero. Thus if t is greater than $4g+p^s+2$ all syndromes in (6) are zero. In order to determine $[\tilde{\omega}^{(2)}(x,y)]$ we see after inspecting (7) that we need to know syndromes $\tilde{S}_{a,b}$ corresponding to pole order $\nu_\infty(\sigma(x,y)) + (p^s-1)(p^r+1) - p^s$. This implies that we need to determine at most $(p^s-1)(p^r+1) - p^s + p^s + 2 = 2g + p^s + 1$ new syndromes. In order to show that the number of coefficients of $\tilde{\omega}(x,y)$ is at most $g+p^s-1$ we assume the existence of a non-zero term $\tilde{\omega}_{a,b}$. At least one term in expression (7) must be nonzero for fixed a, b which leads to the two inequalities

$$lp^s + (p^k - 1 - b + j)(p^r+1) \geq \nu_\infty(\sigma) - p^s - 1 \text{ for } k \leq s$$

$$(a+l+1)p^s + j(p^r+1) \leq \nu_\infty(\sigma)$$

Working out these two inequalities we find

$$ap^s + b(p^r+1) \leq (p^r+1)(p^k-1) - 1 \leq 2g + p^s - 2$$

Thus $\tilde{\omega}(x,y)$ is an element of $L((2g+p^s-2)P_\infty)$ and has at most $g+p^s-1$ coefficients as a consequence of the Riemann-Roch Theorem. □

We propose the following algorithm for determining error values if the number of errors that actually ocurred is at least $4g+p^s+2$.

Algorithm 1

Input: The error-locator of minimal pole order, syndromes $S_{a,b}$, $ap^s+b(p^r+1) < m$, *and the set* G *provided by a Berlkamp-Massey type algorithm.*

Computation:

1. *Determine from a function* $\lambda^{(j)} \in G$ *the modified syndromes* $\tilde{S}_{a,b}$ *for*

$$\nu_\infty(\sigma) - p^s - 1 \leq ap^s + b(p^r+1) \leq \nu_\infty(\sigma) + (p^s-1)p^r - p^s.$$

2. *Determine* $\tilde{\omega}(x, y)$
3. *For all* P_i *such that* $\lambda^{(j)}(P_i) \neq 0$ *and* $\sigma(P_i) = 0$ *determine the error values as*

$$e_i = \frac{(x^{p^{2r}} - x)\tilde{\omega}(x, y)}{\lambda^{(j)}(x_i, y_i)\sigma(x, y)}\Bigg|_{\substack{x = x_i \\ y = y_i}}.$$

4. *if there are points* P_i *such that* $\lambda^{(j)}(x_i, y_i) = 0$ *and* $\sigma(x_i, y_i) = 0$ *then let* $G \leftarrow G \setminus \lambda^{(i)}$ *and goto 1.*

VI Conclusions

We have outlined two formulas for the determination of error values for codes from a class of maximal curves. The first method can be seen as a generalization of Forney's formula for RS codes. This method is conceptionally simpler than the second method which uses a simplified error-evaluator polynomial together with a polynomial from the G-set of a Berlekamp-Massey type algorithm for AG codes. Using this simplified error-evaluator is advantageous if the number of errors is large compared to the genus of the underlying curve. A complexity analysis of the approaches yields that both methods accomplish the task of finding error-values with asymptotic complexity $O(n^2)$ arithmetic operations in the field of definition of the code.

References

[1] S. Sakata, J. Justesen, Y. Madelung, H. Elbrønd Jensen, and T. Høholdt. Fast decoding of algebraic geometric codes up to the designed minimum distance. IEEE Trans. Inf. Theory, vol. 38, pp. 111-119, Jan. 1992.

[2] R. Kötter, *On ALgebraic Decoding of Algebraic-Geometric and Cyclic Codes,* thesis, Linköping, Sweden, 1996.

[3] D. Leonard, *A generalized Forney formula for AG codes* IEEE Trans. Inf. Theory, vol. 42, no. 4, pp. 1263-1268, July 1996.

[4] J.P. Hansen, H.E. Jensen, and R. Koetter, *Determination of Error Values for AG-Codes and the Forney Formula,* submitted to *IEEE Trans. Inf. Theory.*

[5] M. O'Sullivan, *Decoding of One-Point Codes Using Error Evaluator Polynomials* Proceedings ISIT97 p. 377, Ulm, Germany

[6] R Koetter, *A Formula for the Determination of Error Values for Hermitian Codes* Proceedings ISIT97 p. 383, Ulm, Germany

[7] J.H. van Lint, and G. van der Geer *Introduction to Coding Theory and Algebraic Geometry*, Birkhäuser Verlag, Basel, Boston, Berlin, 1989

[8] A. Garcia, and H. Stichtenoth, *Algebraic Function Fields over Finite Fields with Many Rational Places* IEEE Trans. Inf. Theory, vol. 41, no. 6, pp. 1548-1564, November 1995.

[9] S. Miura and K. Kamiya, *Geometric-Goppa codes on some maximal curves and their minimum distance,* presented at the IEEE Workshop on Information Theory, Susono-shi, Japan, 1993

[10] R.E. Blahut, "Theory and practice of error control codes", Reading, Ma.: Addison-Wesley, 1983.

[11] T. Horiguchi, *High Speed Decoding of BCH Codes Using a New Error Evaluation Algorithm,* Electr. and Comm. in Japan, vol. 72, no. 12, part 3, 1989.

Computationally Efficient Blind Equalization

D.R. Brown, P.B. Schniter, and C.R. Johnson, Jr.*

Abstract: Adaptive blind equalization has gained wide-spread use in communication receivers that operate without training signals. In particular, the Constant Modulus Algorithm (CMA) [6, 10] has become a favorite of practitioners due to its LMS-like complexity and desirable robustness properties [4, 5, 14]. This paper presents a signed-error version of CMA (SE-CMA) which is motivated by a further reduction in the computation complexity of the equalizer coefficient update. The authors investigate the performance of SE-CMA and establish connections to previously proposed signed blind algorithms. A novel dithered signed blind algorithm (DSE-CMA) is proposed which overcomes significant SE-CMA convergence issues and possesses similar robustness properties to the standard (i.e. unsigned) CMA.

1 Introduction

Digital communication through band-limited or multipath channels results in intersymbol interference (ISI) which, if left uncompensated, can cause unnecessarily high symbol error rates (SER) at the receiver. One popular approach to the mitigation of ISI, termed linear equalization, employs a linear transversal filter at the receiver input. The optimal equalizer achieves the performance goal of minimum SER, however, since SER is a highly nonlinear function of the filter coefficients [13], the minimum-mean-squared-error (MMSE) criterion is commonly used as a close proxy. The least-mean-squares (LMS) algorithm gives a practical adaptive solution to finding the MMSE equalizer when the receiver shares knowledge of a training sequence. Particular circumstances, such as time-varying channels, may force frequent re-transmission of the training sequence, leading to inefficient use of available bandwidth. These drawbacks motivate schemes for *blind* adaptive equalization, where only the statistics of the transmitted signal need be known.

Several Bussgang [2] blind equalization schemes have been proposed over the last 20 years, including those that minimize the Benveniste-Goursat-Ruget [1] and Godard [6] (or constant modulus [10]) criteria. Over the last decade, the Constant Modulus Algorithm (CMA) has established itself as the most popular method for the blind adaptation of linear equalizers. The primary reasons for CMA's success lie in its LMS-like complexity and robustness to the factors present in a typical implementation scenario: additive channel noise, lack of channel disparity, and "short" equalizer lengths. Furthermore, it has been shown that under "ideal" operating conditions, CMA achieves perfect symbol recovery, thus performing as well as any non-blind equalization algorithm.

Though noted for its computational efficiency, CMA may be further simplified by transforming the bulk of its multiplications during the equalizer update into sign operations. The reductions in implementation cost have the potential of making feasible the application of blind equalization to low-cost communication products.

---

*Authors with Cornell University, Ithaca, NY 14853. Supported in part by NSF grant MIP-9509011 and Applied Signal Technology.

Signed versions of the LMS algorithm leave behind a valuable body of literature (see for example [3, 11]). Immediately evident from this literature (and from routine simulations) is that the signed algorithm behaves quite differently than the original algorithm under normal circumstances. Hence, the application of typical signed-update strategies to CMA may undermine the robustness which motivates its use in the first place.

This paper shows that although SE-CMA performs well in the "ideal" case, the robustness properties of CMA are compromised in a realistic environment. As a solution to the observed defect, we present a novel version of SE-CMA which preserves much of the behavior of the original CMA. It evolves from the consideration of the sign operation as a one-bit quantization and employs a dither signal made suitable by the averaging inherent to the stochastic gradient descent (SGD) update.

1.1 System Model

In this paper, we restrict our focus to the case of a noiseless communication system with a FIR channel model and a fractionally-spaced equalizer (FSE) [9] with sampling interval $T/2$ (where T is the symbol period). The source symbols are chosen from a finite, zero-mean alphabet. The system may be represented by the multirate model in Figure 1, where s_n is the baud-spaced source symbol at sample index n, $\mathbf{c}$ is the the vector representing the fractionally-spaced channel impulse response, $\mathbf{f}$ is the vector containing the fractionally-spaced equalizer coefficients, and y_n is the baud-spaced equalizer output. The number of coefficients in the channel and equalizer response vectors are N_c and N_f, respectively. The system output may be expressed as $y_n = \mathbf{s}^t(n)\,\mathbf{C}\,\mathbf{f}$ where $\mathbf{s}(n) = [s_n, s_{n-1}, \ldots, s_{n-N_s+1}]^t$ is the length-$N_s = \lfloor (N_c + N_f - 1)/2 \rfloor$ vector of baud-spaced source symbols and $\mathbf{C}$ is the $N_s \times N_f$ decimated channel convolution matrix given by

$$\mathbf{C} = \begin{bmatrix} c_1 & c_0 & & & & & \\ \vdots & \vdots & c_1 & c_0 & & & \\ c_{N_c-1} & c_{N_c-2} & \vdots & \vdots & \ddots & c_1 & c_0 \\ & & c_{N_c-1} & c_{N_c-2} & \ddots & \vdots & \vdots \\ & & & & & c_{N_c-1} & c_{N_c-2} \end{bmatrix}.$$

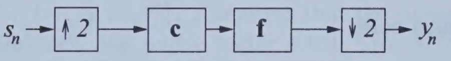


Figure 1: Noiseless $T/2$-spaced multirate system model.

1.2 The Constant Modulus Algorithm

The constant modulus (CM) criterion may be expressed by the non-negative cost function $J_{cma_{p,q}}$ parameterized by positive integers p and q:

$$J_{cma_{p,q}} = \frac{1}{pq} E\{\big||y_n|^p - \gamma\big|^q\}, \qquad (1)$$

where γ is a fixed constant[1] [6, 10]. Two key observations are

---

[1] The CM criterion is most often implemented as CMA 2-2 ($p, q = 2$) and in this case $\gamma = E\{|s_n|^4\}/E\{|s_n|^2\}$. We hereapon assume these values unless noted otherwise.

- The CM cost function is insensitive to phase variations in the received signal. This is useful in receiver designs that attempt to separate carrier synchronization from equalization (e.g. allowing a higher performance carrier recovery loop).
- The CM criterion penalizes the dispersion of received signal modulus, $E\{|y_n|^p\}$, about γ. As a phase-independent criterion for blind source recovery, minimizing dispersion seems intuitively satisfying for CM sources (e.g. M-PSK). Remarkably, the CM criterion also works well with non-CM sources (e.g. M-PAM and M-QAM).

Adaptive implementation using SGD leads to the CMA equalizer update equation:

$$\mathbf{f}(n+1) \;=\; \mathbf{f}(n) + \mu \mathbf{r}^*(n)\psi(y_n), \tag{2}$$

where $\mathbf{f}$ is the length-N_f equalizer coefficient vector, $\mathbf{r}(n)$ is the length-N_f receiver input vector (i.e. regressor), μ is a (small) step-size, and

$$\psi(y_n) \;=\; -\nabla_{y_n} \frac{1}{4}\left(|y_n|^2 - \gamma\right)^2 \;=\; y_n\left(\gamma - |y_n|^2\right) \tag{3}$$

is the CM error function.

It has been shown in [12] that, under ideal conditions, equalizers minimizing the CM cost function perfectly recover the original source symbols, i.e. $y_n = s_{n-\delta}$ for some system delay δ. These perfect blind equalization (PBE) conditions [9] are stated below. Conditions **(A1)** though **(A3)** are required for the possibility of perfect symbol recovery (PSR), whereas conditions **(A4)** and **(A5)** are required to blindly achieve PSR.

(A1) Sufficient equalizer length: For an even-length channel and a $T/2$-spaced FSE, a necessary condition for channel invertibility is that $N_f \geq N_c - 2$.

(A2) Subchannel disparity: For the even-length $T/2$-spaced channel impulse response $\mathbf{c} = [c_0, c_1, \ldots, c_{N_c-1}]$, consider the polynomials formed by the even and odd coefficients of $\mathbf{c}$, i.e. $c_e(z^{-1}) = c_0 + c_2 z^{-1} + \cdots + c_{N_c-2} z^{-\frac{N_c}{2}+1}$ and $c_o(z^{-1}) = c_1 + c_3 z^{-1} + \cdots + c_{N_c-1} z^{-\frac{N_c}{2}+1}$. A necessary condition for perfect channel invertibility is that $c_e(z^{-1})$ and $c_o(z^{-1})$ have no common roots.

(A3) No additive channel noise.

(A4) Sub-Gaussian source: The source kurtosis $\kappa = \frac{E\{|s_n|^4\}}{E\{|s_n|^2\}^2}$ must satisfy $\kappa < 3$ for a real-valued source or $\kappa < 2$ for a complex-valued source.

(A5) White source: The source symbols must be temporally uncorrelated.

An attractive feature of CM equalizers is their robustness to violations of the PBE conditions. For example, CM equalizers have been shown to be quite robust to undermodelling [4], additive channel noise [5, 14], and lack of channel disparity [5].

2 SE-CMA for a Real-Valued Source

Typically, a signed-error algorithm modifies the equalizer update equation of the unsigned algorithm by retaining only the sign of the error function. The attractiveness of this modification is that, when μ is implemented as a fixed bit shift, the N_f regressor multiplies in (2) can be eliminated. Using the standard definition of the sign function for a real-valued argument, (2) can be modified to yield the SE-CMA equalizer update equation

$$\mathbf{f}(n+1) \;=\; \mathbf{f}(n) + \mu \mathbf{r}(n) \underbrace{\operatorname{sgn}\big(\psi(y_n)\big)}_{\triangleq \sigma(y_n)}. \tag{4}$$

The CMA and SE-CMA error functions are compared in Figure 2.

The SE-CMA error function (like CMA) satisfies the Benveniste-Goursat-Ruget Theorem [1] which requires that the function $g(y) = y + \sigma(y)$ satisfies $g(0^+) \geq 0$ and $g''(y) \leq 0$ for $y > 0$, where at least one of the two inequalities must be strict. Since $\sigma(0^+) = 1$ and $\sigma''(y) = 0$ for $y > 0$, we conclude from the BGR Theorem that, under the PBE conditions, SE-CMA converges to an equalizer solution providing perfect symbol recovery.

2.1 Relationships to Other Blind Equalization Algorithms

Consider the CMA 1-1 cost function $J_{cma_{1,1}} = E\{\big||y_n| - \beta\big|\}$. The corresponding SGD algorithm is obtained by taking the instantaneous gradient of $J_{cma_{1,1}}$, given by

$$\begin{aligned} \nabla_{y_n} \big||y_n| - \beta\big| &= \operatorname{sgn}(\beta - |y_n|)\operatorname{sgn}(y_n) \\ &= \operatorname{sgn}\big(y_n(\beta - |y_n|)\big) \end{aligned}$$

for $y \notin \{-\beta, 0, \beta\}$. The CMA 1-1 update then becomes $\mathbf{f}(n+1) = \mathbf{f}(n) + \mu \mathbf{r}(n)\operatorname{sgn}(y_n(\beta - |y_n|))$ which is equivalent to SE-CMA when $\beta = \sqrt{\gamma}$.

Now consider the signed-Sato algorithm [15] (for real-valued observations): $\mathbf{f}(n+1) = \mathbf{f}(n) + \mu \mathbf{r}(n)\operatorname{sgn}(\rho \operatorname{sgn}(y_n) - y_n)$. Using straightforward algebra, it can be shown that the signed-Sato algorithm is equivalent to SE-CMA when $\rho = \sqrt{\gamma}$.

It is also straightforward to show that the Signed Godard Algorithm proposed by Weerackody (WSGA) in [16] with real valued observations is equivalent to SE-CMA.

Thus *SE-CMA* $\Leftrightarrow$ *signed-Sato* $\Leftrightarrow$ *CMA 1-1* $\Leftrightarrow$ *SGA* for real-valued communication systems. The following examples and analyses apply to each of these algorithms.

2.2 Properties of the Noiseless SE-CMA Cost Surface

We denote the set of points in the domain of the SE-CMA error function $\sigma(\cdot)$ leading to discontinuity as $\mathcal{D} = \{-\sqrt{\gamma}, 0, \sqrt{\gamma}\}$. Assuming that the source symbols are selected from a finite M-ary alphabet, the source vector $\mathbf{s}$ will have M^{N_s} possible realizations. We denote this set of all possible $\mathbf{s}$ by $\mathcal{S}$. Using a similar approach as in [3], note that the sizes of $\mathcal{D}$ and $\mathcal{S}$ imply that there are $3M^{N_s}$ hyperplanes of dimension $N_f - 1$ in the equalizer parameter space containing $\mathbf{f}$ that satisfy

$$y = \mathbf{s}^t \mathbf{C} \mathbf{f}, \text{ for } \mathbf{s} \in \mathcal{S} \text{ and } y \in \mathcal{D}. \tag{5}$$

These equalizer-space hyperplanes define boundaries across which the sign of the SE-CMA error function changes. In other words, within any convex region $\mathcal{R} \in \mathbb{R}^{N_f}$ *not* subdivided by any sign boundary, the SE-CMA error function remains constant.

For adequately small step-sizes and equiprobable $\mathbf{s} \in \mathcal{S}$, the expected trajectory step for any equalizers within the interior of a given region $\mathcal{R}$ is

$$E\{\mathbf{f}(n+1) - \mathbf{f}(n)\} = \mu M^{-N_s} \sum_{\mathbf{s} \in \mathcal{S}} \mathbf{s}^t \mathbf{C}\, \sigma(\mathbf{s}^t \mathbf{C} \mathbf{f}_{\mathcal{R}}) \tag{6}$$

where $\mathbf{f}_{\mathcal{R}}$ is any equalizer in $\mathcal{R}$. Observe that there is only one average trajectory associated with each region $\mathcal{R}$.

2.3 Illustrative Examples

In this section, we illustrate several important properties of SE-CMA using examples constructed with two-tap equalizers. This permits visualization of the cost surface in equalizer-coefficient space. Conveniently, the sign-boundary hyperplanes become lines and the regions $\mathcal{R}$ created by these lines are 2-dimensional tiles in the equalizer plane. Furthermore, the fact that the average trajectory update is constant within a tile implies that the two-tap SE-CMA (or CMA 1-1) cost surface is composed of flat facets.

2.3.1 Perfect Blind Equalization with BPSK source

Consider the communication system in Figure 1 with an i.i.d. BPSK source, the noiseless channel $\mathbf{c}_a = [0.2, 0.5, 1, -0.1]^t$, and a two-tap FSE. Here, the set $\mathcal{S}$ is composed of the columns of $\mathbf{S} = \left[\begin{smallmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{smallmatrix}\right]$ and the channel convolution matrix is $\mathbf{C} = \left[\begin{smallmatrix} 0.5 & 0.2 \\ -0.1 & 1 \end{smallmatrix}\right]$. Figure 3 shows the sign-boundaries from (5) overlaid on the CMA 2-2 cost surface contours in equalizer space and several SE-CMA equalizer trajectories. Note the following:

- Within the interior of a given facet, the SE-CMA trajectories tend to follow a linear path in the direction of steepest descent.
- The trajectories closely follow boundaries between two facets whose gradients both point in the direction of the boundary. Looking closely, the trajectories actually "bounce" between such facet pairs.
- The trajectories converge to the intersection of four facets whose gradients all point in the direction of the intersection. These stable intersections coincide with the stable stationary points of the CMA 2-2 cost surface, which in the PBE case correspond to perfect symbol recovery. Given the BGR Theorem, this global convergence is expected. Note that the unstable intersections coincide with the unstable stationary points of the CMA 2-2 cost surface (in this case, 4 saddle points and one maximum).

2.3.2 Perfect Blind Equalization with 8-PAM source

Now consider the same communication system except with source symbols chosen from a unit-variance 8-PAM alphabet. The increase in the size of the source alphabet leads to an exponential increase in the number of sign-boundaries. For a two-tap equalizer, they are plotted in Figure 4.

Notice that the 8-PAM SE-CMA cost surface is considerably more complicated due to the large number of facets. In general, neighboring facets do not necessarily have opposing gradients and thus equalizer trajectories may cross sign boundaries and assume a new average trajectory toward the perfect equalization solution. Also notice that the stable intersections of sign-boundaries do *not* coincide with the PSR minima as expected from the BGR Theorem but instead appear to be collinear with the PSR minima. This is due to the fact that the choice of $\gamma = E\{s^4\}/E\{s^2\}$ as specified by CMA 2-2 is not appropriate for SE-CMA with non-CM sources such as BPSK. However, as the BGR Theorem predicts, there is a value for γ for which SE-CMA converges to the perfect equalizer. Specifically, γ should be chosen such that $\gamma = a_\nu^2$ where a_ν is the ν^{th} positive member of the source alphabet and integer ν satisfies

$$\nu = \arg\min_k \frac{\left(k - \frac{1}{2}\left(1 + \sqrt{\frac{M^2}{2} - 1}\right)\right)^2}{k - \frac{1}{2}}. \tag{7}$$

The proof is provided in Appendix A. For the 8-PAM unit-variance alphabet, (7) suggests $\gamma = 25/21$. Figure 5 shows that perfect equalization is achieved for this value of γ.

2.3.3 Undermodelled channel with BPSK source

The question remains, "How well does SE-CMA work under a practical context?" The following examples illustrate the behavior of SE-CMA in more realistic conditions by violating PBE condition **(A1)**, otherwise known as channel undermodelling. This situation nearly always occurs in practice. Consider the communication system in Section 2.3.1 but with channel $\mathbf{c}_b = [0.1, 0.3, 1, -0.1, 0.5, 0.2]^t$. Figure 6 shows the CMA 2-2 cost surface contours and the SE-CMA sign-boundaries with $\mathcal{S}$ composed of the columns of $\mathbf{S} = \begin{bmatrix} -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}$ and the channel convolution matrix is $\mathbf{C} = \begin{bmatrix} 0.3 & 0.1 \\ -0.1 & 1 \\ 0.2 & 0.5 \end{bmatrix}$.

Examine the two SE-CMA trajectories in Figure 6. The trajectory labeled A converges to a stable intersection point of the sign-boundaries in the neighborhood of (but not collinear with) the CMA 2-2 and MMSE minima. Trajectory B, on the other hand, does not converge to any point on the cost surface after 10000 iterations. This failure to converge can be explained by calculating the average trajectory in this tile using (6), from which we obtain the alarming result that $E\{\mathbf{f}(n+1) - \mathbf{f}(n)\} = 0$. Although the instantaneous trajectory will be non-zero, the zero-valued average update has the potential of capturing the trajectory for an arbitrarily long time.

Simulations suggest that *zero-slope facets* always exist for both BPSK and M-PAM sources in the case of channel undermodelling. This seriously undermines the applicability of SE-CMA to practical communication systems in high-SNR environments. As an alternative, we propose a novel signed-update algorithm that retains the computational efficiency of SE-CMA while exhibiting the desirable robustness properties of CMA 2-2. As such, it is unhindered by the existence of zero-slope facets.

3 Dithered SE-CMA

"Gimme noise, noise, noise, noise ... " —*Paul Westerberg,* Stink, *1982.*

This section describes a simple modification to SE-CMA that results in an algorithm which, for small step-sizes, behaves remarkably like the original CMA while retaining the computational efficiency of SE-CMA.

Viewing the sign operation in SE-CMA as a one-bit quantizer, it makes sense to consider the application of *dither* [7]. Dithering techniques attempt to preserve information lost in the quantization process by making the quantization noise white and independent of the signal being quantized. Such techniques[2] accomplish these properties by adding a random process (*dither*) to the information signal before quantization.

The dithered SE-CMA (DSE-CMA) algorithm is formed by dithering the SE-CMA error function with d_n:

$$\mathbf{f}(n+1) \;=\; \mathbf{f}(n) - \mu \mathbf{r}(n) \underbrace{\operatorname{sgn}\big(\psi(y_n) + d_n\big)}_{\triangleq \varphi(y_n)}. \tag{8}$$

In our case, a simple but useful d_n is an i.i.d. sequence uniformly distributed on $(-1, 1]$. Using the theorems in [7], it can be shown that this dither has the desirable property that the quantization noise $\epsilon_n = \varphi(y_n) - \psi(y_n)$ is zero-mean and uncorrelated with $\psi(y_n)$ for $|\psi(y_n)| < 1$. In other words, the random function φ is a "clipped" version of the CMA error function ψ:

---

[2] See [8] for an application of dither to LMS-based echo cancellation.

$$E\{\varphi|y\} = \begin{cases} -1 & y : \psi(y) < -1, \\ 1 & y : \psi(y) > 1, \\ \psi & \text{else.} \end{cases} \tag{9}$$

For small μ, the averaging inherent in DSE-CMA sufficiently approximates the expectation in 9. In this case, we can conclude that DSE-CMA should closely follow the average CMA 2-2 trajectories. Furthermore, the zero-mean property of the quantization noise leads us to believe that DSE-CMA should exhibit mean convergence to a (local) CMA solution. Figure 7 presents simulation results demonstrating these claims. Note that DSE-CMA trajectories appear to descend the CMA 2-2 error surface, not the faceted SE-CMA error surface. Unlike SE-CMA, the trajectories are not captured by zero-slope facets.

The principle disadvantage of DSE-CMA over CMA 2-2 is its higher *excess CM-cost*, defined as the increase in steady-state CM-cost over the (locally) minimum CM-cost [9]. A variable step-size, or "gearshift" algorithm could help remedy this undesired steady-state behavior without compromising convergence speed.

4 SE-CMA for a complex-valued source

As mentioned in Section 1.2, a key feature of CMA is that its convergence is unhindered by variations in the complex-valued received signal phase. We desire a computationally-efficient algorithm (i.e. no regressor multiplies) with the same property.

First, consider extending the real-valued sgn(·) function in (4) to $\text{csgn}(x) = \text{sgn}(\text{Re}\, x) + j\,\text{sgn}(\text{Im}\, x)$. Using straightforward algebra, we obtain the complex-valued SE-CMA update equation $\mathbf{f}(n+1) = \mathbf{f}(n) + \mu \mathbf{r}^*(n)\, \text{csgn}(y_n)\, \text{sgn}(\gamma - |y_n|^2)$, where γ is a source dependent constant. We can compare this to a technique proposed by Weerackody in [16], termed the Signed Godard Algorithm (WSGA) and specified by the update $\mathbf{f}(n+1) = \mathbf{f}(n) + \mu \mathbf{r}^*(n)\, \text{csgn}(y_n)\, \text{sgn}(\beta - |\,\text{Re}\, y_n| - |\,\text{Im}\, y_n|)$. Note that WSGA attempts to force the equalized constellation onto a square rotated 45° in the complex plane, while SE-CMA attempts to force the constellation to a (constant-modulus) circle.

The phase-dependence of the WSGA "modulus-restoral" term, $\text{sgn}(\beta - |\,\text{Re}\, y_n| - |\,\text{Im}\, y_n|)$, makes it susceptible to convergence problems when in the presence of carrier frequency offset. Intuitively, this carrier frequency offset can be considered a rapidly time-varying channel characteristic for which a small step-size algorithm like WSGA may be incapable of tracking. We note that this is not the case for the "modulus-restoral" term in SE-CMA: $\text{sgn}(\gamma - |y_n|^2)$. However, we should not overlook the phase-dependent term $\text{csgn}(y_n)$ present in both WSGA and SE-CMA.

The phase-independence of the CMA 2-2 update equation given by (2) and (3) is made possible by the interaction of the terms $\mathbf{r}^*(n)$ and y_n: they completely cancel the effects of received-signal phase-change on the equalizer update. In SE-CMA, however, the pair $\mathbf{r}^*(n) y_n$ is replaced by $\mathbf{r}^*(n)\, \text{csgn}(y_n)$, which unfortunately does not retain this phase-independence property. Simulations have shown, however, that the "4-quadrant quantization" imposed by the csgn(·) operation is sufficiently averaged by the recursive nature of SE-CMA, permitting fast convergence in the presence of large or small amounts of carrier-frequency offset.

5 Conclusion

This paper studies the SE-CMA real-valued equalization algorithm (and equivalents), derives the appropriate SE-CMA dispersion constant for M-PAM sources, and demonstrates significant convergence issues in the undermodelled, noiseless case. We propose a dithering technique that addresses this issue while retaining the computational efficiency of SE-CMA. Subsequent research focuses on two open problems: an understanding of the transient and asymptotic behavior of these algorithms in the presence of AWGN and the search for a computationally efficient complex-valued equalization algorithm retaining the phase-insensitivity of CMA.

References

[1] A. Benveniste, et al., "Robust Identification of a Nonminimum Phase System: Blind Adjustment of a Linear Equalizer in Data Communications," *IEEE Trans. on Automatic Control*, vol. 25 no. 3, pp. 385-399, June 1980.

[2] S. Bellini, "Bussgang techniques for blind deconvolution and equalization," in *Blind Deconvolution*, Englewood Cliffs, NJ: Prentice-Hall, 1994, pp. 60-120.

[3] C. R. Elevitch et al., "Quiver Diagrams and Signed Adaptive Filters," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, no. 2, pp. 227-235, Feb. 1989.

[4] T. J. Endres, et al., "On the Robustness of the Fractionally-Spaced Constant Modulus Criterion to Channel Order Undermodeling: Part I," in *Proc. IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications* (Paris, France), pp. 37-40, 16-18 Apr. 1997.

[5] I. Fijalkow, et. al, "Fractionally spaced equalization using CMA: robustness to channel noise and lack of disparity," *IEEE Trans. on Signal Processing*, vol. 45, no. 1, pp. 56-66, Jan. 1997.

[6] D.N. Godard, "Self-Recovering Equalization and Carrier Tracking in Two-Dimensional Data Communication Systems," *IEEE Trans. on Communications*, vol. 28, no. 11, pp. 1867-1875, Nov. 1980.

[7] R.M. Gray et al., "Dithered Quantizers," *IEEE Trans. on Information Theory*, vol. 39, no. 3, pp. 805-812, May 1993.

[8] M. Bonnet and O. Macchi, "A low complexity echo canceller performing the channel phase acquisition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 200-202, Feb. 1986.

[9] C.R. Johnson et al., "Blind Equalization Using the Constant Modulus Criterion: A Review," Submitted to *Proceedings of the IEEE*, Mar. 1997.

[10] J.R. Treichler et al., "A new approach to multipath correction of constant modulus signals," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-31, no.2, pp. 459-72, Apr. 1983.

[11] W. A. Sethares, et al., "Adaptive Algorithms with Nonlinear Data and Error Functions," *IEEE Trans. on Signal Processing*, vol. 40, no. 9, pp. 2199-206, Sept. 1992.

[12] Ye Li et al., "Global convergence of fractionally-spaced Godard (CMA) adaptive equalizers," *IEEE Trans. on Signal Processing*, vol. 44, no.4, pp. 818-26, Apr. 1996.

[13] J.G. Proakis, *Digital Communications*, 3rd ed., New York, NY:McGraw-Hill, 1989.

[14] H.H. Zeng et al., "On the performance of CMA in the presence of noise," in *Proc. Conference on Information Science and Systems* (Princeton, NJ), pp. 890-894, Mar., 1996.

[15] V. Weerackody, et al., "A convergence model for the analysis of some blind equalization algorithms," in *Proc. IEEE International Symposium on Circuits and Systems* (), pp. 2136-2139, 1989.

[16] V. Weerackody, et al., "A simple hard-limited adaptive algorithm for blind equalization," *IEEE Trans. on Circuits and Systems*, vol. 39, no. 7, pp. 482-487, July 1992.

A Selection of γ for M-PAM sources

In this section we derive the unique positive γ that provides perfect blind equalization for SE-CMA and an M-PAM source under the PBE conditions. The BGR Theorem tells us that such a γ exists, and the equivalence between the SE-CMA solutions and the CMA 1-1 minima provides a straightforward method of finding this optimal γ.

We begin with the CMA 1-1 cost function $J(\gamma, \mathbf{h}) = E\left\{\left|\left|\mathbf{s}^t\mathbf{h}\right| - \sqrt{\gamma}\right|\right\}$, written in terms of γ and the system response $\mathbf{h} = \mathbf{Cf}$. Perfect symbol recovery requires that $\mathbf{h} = \mathbf{e}_\delta$, where $\mathbf{e}_\delta$ denotes a standard basis vector (i.e. $e_\delta = 1$ and $e_{i \neq \delta} = 0$). Before proceeding, we make two important observations: (i) the BGR Theorem implies that $J(\gamma_{opt}, \mathbf{h})$ is minimized by $\mathbf{h} = \mathbf{e}_\delta$, and (ii) the form of $J(\gamma, \mathbf{h})$ implies that changes in γ scale the cost surface radially in $\mathbf{h}$-space. From these points we conclude that, for any γ, the CMA 1-1 minima are constrained to the set $\alpha\, \mathbf{e}_\delta$ (where α is a real scalar constant). This allows us to define the problem as follows:

$$\text{Find } \gamma \text{ such that } J(\gamma, \mathbf{e}_\delta) = \min_{\alpha} J(\gamma, \alpha\mathbf{e}_\delta). \tag{10}$$

For an M-PAM equiprobable source, $J(\gamma, \alpha\mathbf{e}_\delta) = \frac{2}{M}\sum_{l=1}^{M/2}|\alpha a_l - \sqrt{\gamma}|$, where a_l denotes the l^{th} positive alphabet member. Using straightforward algebra, it can be shown that the value of α minimizing this quantity is $\alpha_{min} = \sqrt{\gamma}/a_\nu$ for the integer ν satisfying

$$\nu = \arg\min_{k} \sqrt{\gamma}\left(\frac{\left(\frac{M}{2}\right)^2 + 2k^2 - 1}{2k-1} - \left(\frac{M}{2} + 1\right)\right) \tag{11}$$

Completing the square in (11) yields (7). Since ν may be determined without knowledge of γ, we can rewrite (10) as

$$\text{Find } \gamma \text{ such that } \frac{2}{M}\sum_{l=1}^{M/2}|a_l - \sqrt{\gamma}| = \frac{2}{M}\sum_{l=1}^{M/2}\left|\frac{\sqrt{\gamma}}{a_\nu}a_l - \sqrt{\gamma}\right|, \tag{12}$$

from which it is clear that $\gamma_{opt} = a_\nu^2$.

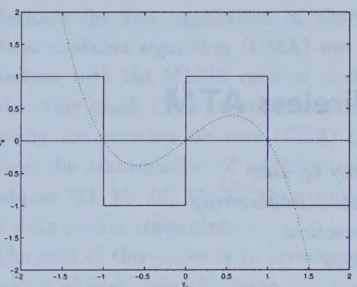


Figure 2: SE-CMA error function $\sigma(\cdot)$ (solid), and CMA error function $\psi(\cdot)$ (dotted), for $\gamma = 1$.

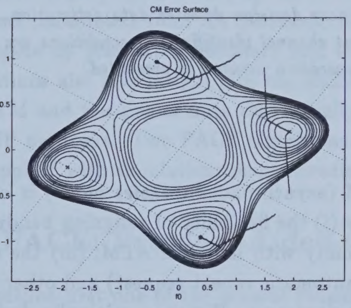


Figure 3: SE-CMA trajectories and sign-boundaries overlaid on CMA 2-2 cost contours for BPSK over channel $\mathbf{c}_a$.

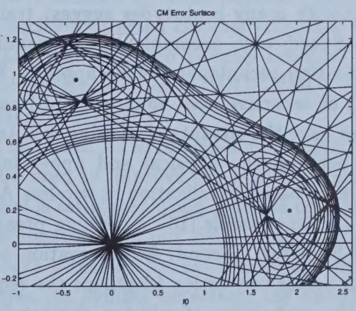


Figure 4: SE-CMA sign-boundaries for 8-PAM over noiseless channel $\mathbf{c}_a$ with γ chosen as for CMA 2-2.

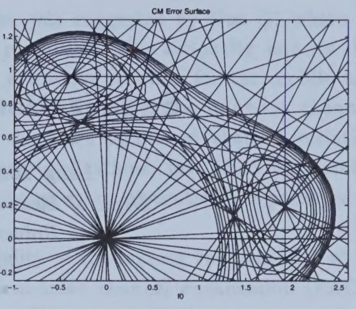


Figure 5: SE-CMA sign-boundaries for 8-PAM over noiseless channel $\mathbf{c}_a$ with γ chosen via Section 2.3.2.

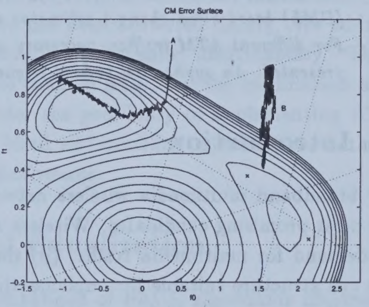


Figure 6: SE-CMA trajectories and sign-boundaries for BPSK over noiseless channel $\mathbf{c}_b$ with $\gamma = 1$.

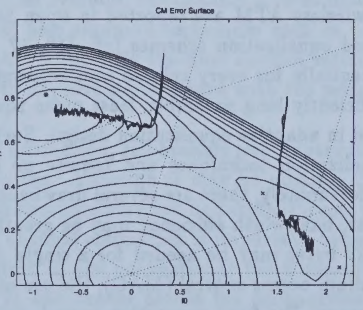


Figure 7: (Rough) DSE-CMA trajectories overlaid on (smooth) average CMA trajectories for BPSK over noiseless channel $\mathbf{c}_b$.

Equalizations in Wireless ATM

Lang Tong and Jeffrey Q. Bao
Dept. of Electrical and Systems Engineering
University of Connecticut
Storrs, CT 06269

Abstract

Equalization in wireless ATM without using preamble is considered in this paper. We propose a protocol-aided channel equalization (PACE) method for wireless ATM network with time- and frequency-selective channels. Exploiting the structure of the ATM cell and the medium access control (MAC), PACE includes a deterministic maximum likelihood (DML) based blind channel estimator and a nonlinear decision directed data estimation. For different ATM traffics, necessary and sufficient channel identifiability conditions are presented. An application to high frequency communication network is presented.

1 Introduction

For broadband multimedia wireless networks, ATM (asynchronous transfer mode) is one of the most promising candidates. Wireless ATM offers (i) the flexibility of assigning bandwidth on demand for multimedia traffic; (ii) the compatibility with land-line ATM; (iii) the small size cell structure suitable for implementing ARQ (automatic repeat request) and other error protection measures which are critical in wireless environment. However, originally designed for networks with high quality physical links, ATM in wireless network requires sophisticated signal processing to combat time varying multipath fading. Because of the high data rate of ATM, TDMA wireless channels are highly time-dispersive. One of the key components in wireless ATM is channel equalization [2].

While the need of equalization in wireless ATM is well known, few details about equalization in wireless ATM are reported in open literature although many prototypes suggest training based equalization schemes (cf. [18, 17, 2, 10, 9]). Specifically, preambles are transmitted, potentially for every cell, so that conventional equalization techniques can be applied. With sufficiently long preamble, least mean squares (LMS) or recursive least squares (RLS) can be used in adaptive equalization design. Such approaches have the advantage of using well tested equalization techniques and require no interaction between the physical and upper layers. Unfortunately, there are several drawbacks of the training based equalization in wireless ATM. First, bandwidth efficiency may be greatly affected [2]. For example [18], a training preamble of 27 bytes may be needed for a single ATM cell of 53 bytes. Second, for reservation-based medium access control (MAC), the use of preamble in access contention lengthens the request message. For fixed contention period, this increases the probability of collision, further limits network efficiency. To improve the overall efficiency in bandwidth utilization, it is necessary to reduce or even eliminate the overhead associated with the training. Potentially this can be achieved by the use of *blind* equalization methods [21, 8].

Perhaps the first application of blind equalization was suggested by Cox [7] where the constant modulus algorithm (CMA) was considered promising. The simplicity and its close connection with the MMSE receiver make CMA an attractive solution for the mobile user. On the other hand, CMA lacks the convergence speed for the equalization of uplink channels, especially for constant bit rate (CBR) traffic when there are significant channel variations between the transmission of adjacent cells. For these applications, fast blind equalization techniques [23, 25, 16, 13, 21] that exploit both time and space diversities are preferred for short data packet transmissions.

The goal of this paper is to investigate the possibility of eliminating the preamble at the beginning of each burst by using the so-called protocal-aided channel equalization (PACE). This is motivated by the fact that, in wireless ATM, cell headers are often transmitted and many MAC protocols [18, 17, 15] are designed, or may be modified, in such a way that headers are known *a priori* to the receiver. The key component of PACE is blind channel estimation that must have fast convergence rate, possiblely achieving equalization within one data cell. The requirement of convergence speed favors maximum likelihood (ML) based estimators such as IQML [6, 20] and TSML [13]. However, fundamental questions about the identifiability of channel parameters from the likelihood function have not been answered clearly. We address this issue by presenting necessary and sufficient conditions of channel identifiability. These conditions also give insight into the limitation on the channel length by the modulation used in ATM and the size of data cell. Finally, we consider the possibility of implementing ATM for HF networks. Two PACE algorithms based on blind and nonblind equalization methods are applied with simulations demonstrating their effectiveness.

2 PACE: Protocol-Aided Channel Equalization

We outline first one of the several possible implementations of PACE algorithms illustrated in Figure 1. Details can be found in [4]. Similar to [26], our approach is a combination of (blind) channel estimation and the nonlinear data directed estimator (NDDE) [11] although other block detection approaches [12] can be easily incorporated. Given the received data from one or a block of cells, an initial channel estimate $\tilde{\mathbf{h}}$ is obtained by a maximum likelihood based blind channel estimation. The initial channel estimate is then used in NDDE to produce a tentative decision $\tilde{\mathbf{s}}_p$ of the payload. Using these detected symbols, the channel is re-estimated by a decision directed channel estimator (DDCE) that uses standard techniques such as LMS or RLS. The improved channel estimate $\hat{\mathbf{h}}$ is used next to provide the final symbol estimate $\hat{\mathbf{s}}_p$. Obviously additional DDCE-NDDE can be used to improve performance with the price of increased complexity.

The role of ATM cell header is twofold. In channel estimation, cell headers and its structure can be used to improve the quality of channel estimation. Although it may be possible that channel is estimated using only the cell header, the performance may not be satisfactory as will be demonstrated in our simulation. The use of blind channel estimation is important because it blocks possible channel estimation error propagation in the conventional block detection schemes. In symbol detection, the knowledge of ATM cell header blocks the propagation of detection errors that severely hinders the detection performance.

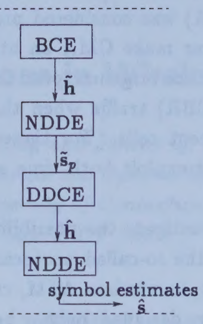


Figure 1: PACE flow diagrams

3 Deterministic ML Estimation and Channel Identifiability

We now focus on the blind channel estimation part of PACE, especially on the issue of when deterministic ML based approaches can be applied.

3.1 ATM Channel Models

The baseband model of the physical link is given by

$$y(t) = \sum_k s_k h(t - kT) + n(t), \tag{1}$$

where $h(t - kT)$ is the composite baseband channel. Here we have assumed that, for ATM applications where users are not traveling in high speed, that the channel is time invariant within one or a few cells. Depending on the radio propagation environment, the baseband channel may last from 10 to as much as 100 symbol intervals [2, 1]. When $y(t)$ is fractionally sampled, say at the rate twice of the symbol rate, we then have the equivalent discrete-time multichannel model ($P = 2$) shown in Figure 2 where $h^{(i)}$ are the even and odd samples of the channel. Of course, if antenna array is used, the same model holds for $P > 2$.

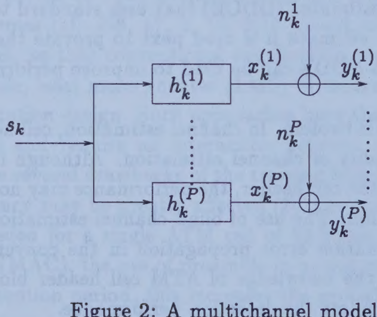


Figure 2: A multichannel model.

Denoting the vector impulse response by

$$\mathbf{h}_k \triangleq [h_k^{(1)}, \cdots, h_k^{(P)}]^t, \tag{2}$$

we have the following system equations in the time domain

$$\mathbf{x}_k = \sum_{i=0}^{L} \mathbf{h}_i s_{k-i} + \mathbf{n}_k, \tag{3}$$

where $\mathbf{y}_k$ is the received (noisy) signal. Depending on the type of traffic and the mode of transmission, the channel model takes slightly different forms.

Block Transmission Model

For up link CBR or VBR traffic, each user transmits at predetermined time slots a block of N cells. In such a case, guard period is included in order to separate different users. Let $\mathbf{x}(t)$ be the vector containing all data transmitted during the time-slot t, we have

$$\mathbf{x}(t) = \mathcal{F}(\mathbf{h})\mathbf{s}(t) + \mathbf{n}(t), \quad t = 1, \cdots, N, \tag{4}$$

where

$$\mathcal{F}(\mathbf{h}) = \begin{pmatrix} \mathbf{h}_0 & & & \\ \vdots & \ddots & & \\ \mathbf{h}_L & & \mathbf{h}_0 & \\ & \ddots & & \ddots \\ & & \mathbf{h}_L & \mathbf{h}_0 \\ & & & \ddots & \vdots \\ & & & & \mathbf{h}_L \end{pmatrix} \tag{5}$$

is the filtering transform assuming the existence of guard-time before and after the transmission, $\mathbf{s}(t)$ is the vector of transmitted symbols, and $\mathbf{n}(t)$ is the additive noise. Note that t is the time index of the data block. In general, data blocks in slot t_1 and slot t_2 are unrelated. It is often the case that channel equalization must be performed within the time slot, i.e., $N = 1$.

Sub-Block Transmission Model

This model is the same as the standard single channel communication problem, and it is appropriate for the downlink or ABR traffic in the uplink. It can also be applied to CBR and VBR traffics in the uplink provided that the data block is sufficiently long. Consider a vector $\mathbf{x}_q(t)$ of q consecutive received samples, we have

$$\mathbf{x}_q(t) = \mathcal{H}_q(\mathbf{h})\mathbf{s}(t) + \mathbf{n}(t), \quad t = 1, \cdots, N, \tag{6}$$

where

$$\mathcal{H}_q(\mathbf{h}) = \underbrace{\begin{pmatrix} \mathbf{h}_L & \cdots & \mathbf{h}_0 & & \\ & \ddots & & \ddots & \\ & & \mathbf{h}_L & \cdots & \mathbf{h}_0 \end{pmatrix}}_{L+q \text{ columns}} \tag{7}$$

is the filtering transform without guard-time, $\mathbf{s}(t)$ is the vector of $L + q$ consecutive input symbols, and $\mathbf{n}(t)$ is the additive noise.

The sub-block model given above is different from the block transmission in two aspects. First, $\mathbf{x}(t)$ and $\mathbf{x}(t+1)$, similarly, $\mathbf{s}(t)$ and $\mathbf{s}(t+1)$, are related by a shifting operation, and it is often possible to use a number of sub-blocks in channel estimation and equalization. Second, the filtering matrix $\mathcal{H}_q(\mathbf{h})$ may very well be singular depending on the value of q and the nature of the channel $\mathbf{h}$. The full rank condition of $\mathcal{H}_q(\mathbf{h})$ is necessary for channel identification.

3.2 Deterministic ML and Identifiability

Consider the block or the sub-block transmission models in (4- 6) with deterministic input $\mathbf{s}(t)$ and zero-mean i.i.d. Gaussian noise $\mathbf{n}(t)$. The (deterministic) maximum likelihood (DML) estimation of the channel $\mathbf{h}$ and the input $\mathbf{S} = [\mathbf{s}(1), \cdots \mathbf{s}(N)]$ is given by

$$\{\hat{\mathbf{h}}_{ML}, \hat{\mathbf{S}}_{ML}\} = \arg\max_{\mathbf{h},\mathbf{S}} f(\mathbf{X}; \mathbf{h}, \mathbf{S}), \tag{8}$$

where $f(\cdot)$ is the likelihood function, $\mathbf{X} \overset{\Delta}{=} [\mathbf{x}(1), \cdots \mathbf{x}(N)]$ is the observation. Several algorithms, most notablely the TSML approach by Hua [13] and the variation of IQML [6] by Slock [20], have been proposed for the sub-block transmission model, and they are equally applicable for the block transmission.

Considered in this section is the issue of channel identifiability: *can* $\theta = (\mathbf{h}, \mathbf{S})$ *be uniquely determined from the likelihood function* $f(\mathbf{X}; \theta)$*?* Under the (deterministic) maximum likelihood formulation of channel estimation, channel identifiability is defined following [19].

Definition 1 *Let* $\mathbf{x}$ *be a random vector with pdf* $f(\mathbf{x}; \theta) \in \mathcal{F} = \{f(\mathbf{x}; \theta), \theta \in \Theta\}$*. Parameters* θ_1 *and* θ_2 *are observationally equivalent if*

$$f(\mathbf{x}; \theta_1) = f(\mathbf{x}; \theta_2), \quad a.e. \tag{9}$$

Then θ *is said to be identifiable (by the likelihood function) if there is no other parameter* $\theta' \in \Theta$ *that is observationally equivalent.*

The main goal is to characterize Θ, not in a contrived manner, but in a meaningful way that the identifiability condition can be tested.

Applying to the channel identification problems considered in this paper, the above classical definition of identifiability has two equivalent forms.

Theorem 1 *For the both the block and sub-block transmission models, the following statements are equivalent:*

1. $\theta = (\mathbf{h}, \mathbf{S})$ *is identifiable.*

2. $(\mathbf{h}, \mathbf{S}) \neq (\hat{\mathbf{h}}, \hat{\mathbf{S}}) \Leftrightarrow \mathcal{F}(\mathbf{h})\mathbf{S} \neq \mathcal{F}(\hat{\mathbf{h}})\hat{\mathbf{S}}$. *(The same statement applies to the sub-block transmission model* $\mathcal{H}_q(\mathbf{h})$*.)*

3. *The Fisher information matrix is rank deficent by one.*

The last two forms of identifiability were referred to as strictly identifiability and Fisher Information Identifiability by Hua [14] and was shown to be equivalent for the sub-block transmission model.

3.3 Identifiability for Sub-Block Transmissions

Consider first the sub-block transmission model given in (6). The issue of identifiability has been addressed in [25, 13, 14]. Several necessary conditions and sufficient conditions are found separately. The necessary and sufficient condition given in [25] is rather complicated and it is not imposed on the channel and its input separately. We present here a simple necessary and sufficient condition by the use of the concept of linear complexity of input sequence s_k.

Following Blahut [5], the linear complexity of sequence $\{s_k\}_{k=1}^{n}$ is defined as the smallest value of c for which

$$s_i = -\sum_{j=1}^{c} \lambda_j s_{i-j}, \quad i = c, \cdots, n. \tag{10}$$

Define a Toeplitz matrix $\mathbf{S}_c$ by

$$\mathbf{S}_c \triangleq \begin{pmatrix} s_c & s_{c+1} & \cdots & s_n \\ s_{c-1} & s_c & \cdots & s_{n-1} \\ \vdots & \vdots & & \vdots \\ s_1 & s_2 & \cdots & s_{n-c} \end{pmatrix}. \tag{11}$$

If s_k has linear complexity greater than c, $\mathbf{S}_c$ has full row rank. Hence, the sample covariance of the vector sequence $\mathbf{s}(k) = [s_k, s_{k-1}, \cdots, s_{k-c}]^t$ has full rank. On the other hand, if s_k has linear complexity no greater than c, the sample covariance of $\mathbf{s}_k$ is rank deficient.

Theorem 2 *Consider the sub-block transmission model with two sub-channels, i.e., ,* $P = 2$. $\mathbf{h}$ *is identifiable if and only if the signal subspace of the sample covariance*

$$\hat{\mathbf{R}}_x = \frac{1}{N} \sum_t \mathbf{x}_{L+1}(t) \mathbf{x}_{L+1}^H(t) \tag{12}$$

has dimension $2L+1$*, or equivalently,*

1. $\{h_i(z)\}$ *are co-prime;*
2. $\{s_k\}$ *has linear complexity greater than* $2L+1$.

Proof: The sufficiency of the condition can be shown by applying either the least-squares [25] or the subspace [16] blind channel estimators to $\mathbf{x}_{L+1}(k)$. To prove the necessary part, we proceed with two steps. First when $\{h_i(z)\}$ are not co-prime, then $\mathcal{H}_q(\mathbf{h})$ does not have full column rank for all q [22]. It is then clear that there exists $\mathbf{W}$ such that $\mathbf{H}_q(\mathbf{h})\mathbf{S} = \mathbf{H}_q(\mathbf{h})(\mathbf{S} + \mathbf{W})$. We now consider the case when $\{h_i(z)\}$ are co-prime but $\{s_k\}$ has linear complexity no greater than $2L+1$. From the sub-block model, we have

$$\mathbf{X}_{L+1} = \mathcal{H}_{L+1}(\mathbf{h}) \mathbf{S}_{L+1}, \tag{13}$$

where $\mathcal{H}_{L+1}(\mathbf{h})$ has full column rank and $\mathbf{S}_{L+1}$ does not have full row rank. Therefore, there exists a vector $\mathbf{u} \notin \mathcal{R}_{\mathbf{H}_{L+1}^{\perp}}$, hence $\mathbf{u} \neq \mathbf{h}$, such that (i) $\mathbf{u}^H \mathbf{X}_{L+1} = \mathbf{0}$; (ii)$\mathcal{H}_{L+1}(\mathbf{u})$ has full column rank. Here, $\mathcal{R}_{\mathbf{A}}$ and $\mathcal{R}_{\mathbf{A}^{\perp}}$ denote the range space of $\mathbf{A}$ and its orthogonal complement, respectively. Since $\mathcal{R}_{\mathbf{X}} \in \mathcal{R}_{\mathcal{H}_{L+1}(\mathbf{u})}$, we have

$$\mathcal{H}_{L+1}(\mathbf{h}) \mathbf{S}_{L+1} = \mathcal{H}_{L+1}(\mathbf{u}) \mathcal{H}_{L+1}^{\dagger}(\mathbf{u}) \mathbf{X}_{L+1} = \mathcal{H}_{L+1}(\mathbf{u}) \mathbf{W}. \tag{14}$$

It can be verified that $\mathbf{W}$ has the same structure as in $\mathbf{S}$, hence the unidentifiability. □□□

Remarks:

1. Theorem 2 implies that if a channel is to be identified from the observation alone, there must be $4L+1$ symbols in the cell. Given a fixed modulation, say, QPSK, the maximum channel can be identified is 48 taps. When 64-QAM is used, the maximum length allowed is only 4.

2. The concept of signal subspace in the linear model is widely used in array signal processing. The necessary and sufficient condition presented here can be tested by AIC or MDL [24] criteria.
3. The concept of linear complexity related to the notion of "modes" used in [25].

3.4 Identifiability for Block Transmission

Consider now the block transmission model given in (4). The existence of guard-time greatly simplifies the identifiability condition. The following results are easy to obtain in the frequency domain.

Theorem 3 *For block transmission model with N independent blocks,*

N=1: $\mathbf{h}$ *is identifiable if and only if* $\{h_i(z)\}$ *are co-prime and* $\mathcal{Z}(\mathbf{s}(1)) \neq 0$.

N >1: $\mathbf{h}$ *is identifiable if and only if* $\{s_n(z)\}$ *are coprime, where* $s_n(z)$ *is the Z-transform of the nth cell.*

We note that the existence of guard time removes the requirement on the input sequence. When more than two blocks are used, the identifiability is almost always assured.

4 Applications in High Frequency Network

High frequency (HF) band has perhaps one of the most difficult channels for data communication. The rapid time variation is such that as much as 50% of the transmission may be used for channel estimation and equalization. Interestingly, the MIL-STD-188-110A modem standard specifies a channel probe of the size equal to that of an ATM cell. This prompts the design of a HF ATM network protocol that allows the transmission of ATM cell without additional overhead.

The HF ATM network being considered consists of nodes interconnected by the HF link such as the navy tactical HF communication network. The network topology mimics a bus because of the full connectivity of the HF link. The access to the common HF link is coordinated by a Medium Access Control (MAC) protocol resides at every node in a distributed way. The MAC in our simulation is a central controlled Priority Oriented Demand Assigned (PODA) as used in satellite networks. Details of this approach will be reported elsewhere [3].

The Constant Bit Rate (CBR) traffic such as voice communication is difficult for equalization. This is due to the fact that, for HF communication, the channel can not be assumed to be constant from one burst to the next. In our simulation, we considered the worst case when there is only one cell in each burst. The size of a cell is 54 bytes (53 ATM cell plus one byte of possible wireless overhead such as sequence number or CRC) and no preamble is transmitted before each cell although we assume that the link has already been established. Using the Watterson model, we considered a two-ray multipath HF channel with both paths having equal power. Square-root raised cosine filtering was employed at both the transmitter and the receiver sides. The received signal was fractionally sampled at a rate twice of the symbol rate. Because of the possible header compression, we tested the possibility of using headers different length. The simulation parameters are listed in Table 1.

We tested two different approaches using PACE for different header size. One relies purely on the header information for initial channel estimation, i.e. replacing the blind channel

Parameters	General	
Modulation	8PSK	
Symbol rate (Baud)	2400	
SNR(dB)	17-27	
Rolloff factor	0.25	
Length of header (symbol)	0-24	
Length of cell (symbol)	144	
Channel	path #1	path #2
Delay(ms)	0.08	2.08
Doppler shift (Hz)	0	0
Doppler spread (Hz)	0.5	0.5
gain	1	1

Table 1: Simulation Parameters

estimation (BCE) block in Figure 1 by a least squares estimator using header as training symbols. The second approach employs blind channel estimation that uses data for the entire block. For symbol detections, both methods use the header information in NDDE.

Figure 3 shows the performance of the two approaches for header length from 0 to 24. The use of BCE (the solid lines) has advantage over those rely entirely on the knowledge of the header (dashed lines). It is interesting to observe the effects of the header size. Since blind channel estimation does not use explicitly the header information,the symbol detection part does, the performance is relatively insensitive to the header length. In contrast, the nonblind channel estimator is affected greatly by the amount of header information. In our simulation, when the header length is insufficiently long, decisions from previous detections are borrowed. As channel varies, errors in previous detection affect the accuracy of the channel estimates, which in turn affects the detection of new symbols. This is the reason for the poor performance when the header length is at 0 or 8. But when the header is totally known and length becomes sufficient (header=16 and 24), the initial channel estimation using non blind performs reasonably well, so is the over all Symbol Error Rate (SER) performance. Note that, since the channel varies considerablely, the channel estimates from the header are not necessarily accurate for the entire cell. This is one of the reasons that blind channel estimator that uses the entire data block has a clear advantage.

Acknowledgement

Part of the work was supported in part by the National Science Foundation under Contract NCR-9321813, the Office of Naval Research under Contract N00014-96-1-0895 and by the Advanced Research Projects Agency monitored by the Federal Bureau of Investigation under Contract No. J-FBI-94-221. The first author like to thank several useful discussions with Hanks. H. Zeng and Professor Rueywen Liu related to the identifiability conditions.

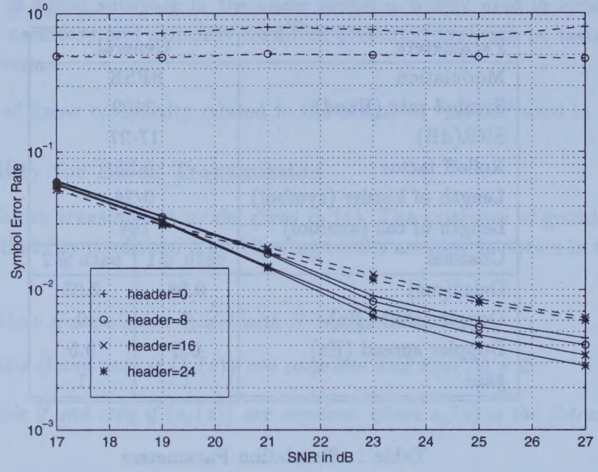


Figure 3: Symbol Error Rate vs. SNR. Solid lines: PACE with blind channel estimation. Dashed lines: PACE with header-aided channel estimation.

References

[1] S. Ariyavisitakul and L. J. Greenstein. "Reduced-complexity equalization techniques for broadband wireless channels". *IEEE J. Selected Areas in Communications*, 15(1):5–15, January 1997.

[2] E. Ayanoglu, Kai Y. Eng, and M. Karol. "Wireless ATM: Limits,Challenges,and Proposals". *IEEE Personal Communications*, pages 18–34, August 1996.

[3] J. Bao and L. Tong. "An Application of Wireless ATM to HF Communication". to be submited.

[4] J. Bao, Y. Zhang, and L. Tong. "Protocol-Aided Channel Equalization in Wireless ATM". May 1998. submited to ICC'98.

[5] R.E. Blahut. *Algebraic Methods for Signal Processing and Communications Coding*. Springer-Verlag, New York, NY, 1992.

[6] Y. Bresler and A. Macovski. "Exact maximum likelihood parameter estimation of supperimposed exponential signals in noise". *IEEE Trans. ASSP*, ASSP-34(10):1081–89, October 1986.

[7] D.C. Cox. "Wireless Personal Communications: What is it?". *IEEE Personal Communications*, 2(1):22–35, Feb 1995.

[8] C.R. Johnson et. al. "Blind equalization using the constant modulus criterion: A review". *IEEE Proceedings*, 1998.

[9] D. Falconer. "A system architecture for broadband milimeter wave access to an ATM Lan". *IEEE Personal Communications*, (8):36–41, August 1996.

[10] L. Fernandes. "Developing a system concept and technologies for mobile broadband communications". *IEEE Personal Communications*, 2(1):54–59, Feb 1995.

[11] F.M.Hsu. "Data directed estimation techniques for single-tone HF modems' ". In *Proc. IEEE MILCOM'85*, pages 12.4.1–12.4.10, 1985.

[12] G.K.Kaleh. "Channel equalization for block transmission systems". *IEEE J. Sel. Areas Comms.*, 13(1):110–121, January 1995.

[13] Y. Hua. "Fast maximum likelihood for blind identification of multiple FIR channels". *IEEE Trans. Signal Processing*, SP-44(3):661–672, March 1996.

[14] Y. Hua and M. Wax. "Strict identifibility of multiple FIR channels driven by an unknown arbitrary sequence". *IEEE Trans. Signal Processing*, SP-44(3):756–759, March 1996.

[15] M. Karol, Z. Liu, and K. Eng. "Distributed Queuing Request Update Multiple Access (DQRUMA) for Wireless Packet (ATM) Networks". In *Proc of International Conference on Communications*, pages 1224–1231, Seattle, USA, June 1995.

[16] E. Moulines, P. Duhamel, J.F. Cardoso, and S. Mayrargue. "Subspace-Methods for the Blind Identification of Multichannel FIR Filters". *IEEE Trans. SP*, SP-43(2):516–525, Feb. 1995.

[17] J. Porter and etal. "The ORL Radio ATM system: Architecture and Implementation". Technical report, Olivetti Research Limitted, January 1996.

[18] D. Raychaudhuri and et. al. "WATMnet: A Prototype Wireless ATM System for Multimedia Personal Communication". *IEEE J. Selected Areas in Communications*, 15(1):83–95, January 1997.

[19] T.J. Rothenberg. "Identification in parametric models". *Econometrica*, 39, 1971.

[20] D. Slock. "Blind Fractionally-Spaced Equalization, Perfect Reconstruction Filterbanks, and Multilinear Prediction". In *Proc. ICASSP'94 Conf.*, Adelaide, Australia, April 1994.

[21] L. Tong and S. Perreau. "Multichannel blind channel estimation: From subspace to maximum likelihood methods". *to appear in IEEE Proceedings, 1998.*

[22] L. Tong, G. Xu, B. Hassibi, and T. Kailath. "Blind identification and equalization of multipath channels: A frequency domain approach". *IEEE Trans. Information Theory*, 41(1):329–334, January 1995.

[23] L. Tong, G. Xu, and T. Kailath. "Blind identification and equalization based on second-order statistics: A time domain approach". *IEEE Trans. Information Theory*, 40(2):340–349, March 1994.

[24] M. Wax and T. Kailath. "Detection of signals by information theoretic criteria". *IEEE Trans. ASSP*, ASSP-33(2):387–392, April 1985.

[25] G. Xu, H. Liu, L. Tong, and T. Kailath. "A Least-Squares Approach to Blind Channel Identification". *IEEE Trans. Signal Processing*, SP-43(12):2982–2993, December 1995.

[26] Q. Zhang and L. Tong. "Applications of Blind Channel Estimation to HF Communication". In *Proc. 1st IEEE Signal Processing Worksho on Signal Processing Advances in Wireless Communication*, volume 1, Paris, France, April 1997.

Coupled Adaptation in Wireless Multiple Access Systems

Gregory J Pottie[1]
Electrical Engineering Dept.
University of California, Los Angeles
405 Hilgard Ave
Los Angeles CA 90095
pottie@icsl.ucla.edu

Abstract

In multiple access wireless systems, adaptive algorithms for power control, antenna arrays, and equalizers can all lead to significant capacity increases. However, independent adaptation of these algorithms is not possible, since changes in beam pattern affect the interference power and thus the required transmitted power level and the residual multipath components, which in turn cause the equalizer to need readjustment. There is a further coupling among users due to the mutual interference. Thus care must be exercised in applying simple gradient descent methods. We describe the nature of the coupling and what modifications must be made to the distributed adaptive methods to ensure stability.

I. Introduction

The variability of wireless channels presents both challenges and opportunities in designing multiple access communications systems. To maximize throughput for a given power budget, the link must adapt to the actual channel conditions, changing the transmitter power level, antenna beam pattern, equalizer settings, and possibly the symbol rate and constellation size. On the other hand, the attenuation and directionality of signals makes possible re-use of the time/frequency resources in space, permitting a large number of users to access the shared medium. These users are coupled by the mutual interference they cause one another. For an acceptable quality of service, each user will typically need a signal to interference ratio (SIR) above some target. The techniques which enhance reliability of a link also affect the interference seen by other users, as do multiple access techniques. Thus, it is counterproductive to separate consideration of the physical and medium access layers in designing a wireless transmission system.

A fundamental limit on the capacity of a multiple access wireless system is imposed by the inability of the various users to perfectly estimate and predict the time-varying channel and interference. In the limit of large complexity and power levels, all other impairments can in principle be overcome if the channel dynamics (including interference) are slow enough. We describe practical techniques for ameliorating the impairments, through channel probing, dynamic power and bit allocation, and adaptive antenna arrays and equalizers. These have been developed in the context of an experimental high speed frequency hopping radio being developed at UCLA. This radio can change the M-QAM constellation on each hop, and has four antenna branches and a decision feedback equalizer. Throughout we pay particular attention to how the channel and interference couple the adaptation of the different techniques, in both desired and undesired fashions.

---

[1] This work was supported in part by DARPA contract JFBI 94-222/J4C942220, and a Motorola University Partnership in Research grant.

In support of the algorithmic development reported here, considerable simulation effort has been involved. For the work which includes antenna arrays, the correlated multipath channel was constructed using ray tracing techniques, which included up to quadruple bounces in a large room [1]. Otherwise, statistical models assuming Rayleigh fading and log-normal shadowing with fourth power distance loss were used. The specific models are relatively unimportant to the conclusions, as we were mainly interested in devising algorithms that would be robust with respect to channel variability, rather than in meeting specific probability of error requirements for the overall system. We presume the reader is familiar with standard diversity techniques, and thus after briefly describing the estimation problem for multiple access systems in section II, in section III we describe the set of adaptive techniques being proposed. In section IV we describe their interactions, and in section V present our conclusions.

II. Estimation of the Channel State

The most effective methods for mitigating channel variability require that the receiver has some knowledge of the time-varying channel impulse response and the mean and variance of the combined noise and interference. We refer to this kind of information as the channel state. There are well-defined limits on how well the channel state can be learned (e.g., the Cramer-Rao bound [2]), even given infinite receiver complexity, and much lower limits if low-complexity iterative techniques are to be used. It is the channel estimation problem that fundamentally limits the performance of digital radios in conditions of fast channel change or low SNR.

In multiple access systems, both the channel and interference are time-varying, and we must estimate both in forming an antenna beam or adjusting an equalizer. If we could perfectly estimate the angles of arrival of the interference (including multipath), with a large enough antenna array we could completely null out the interference, and thus achieve extremely high bandwidth re-use. The time variations in the channel and especially the interference place a limit on the accuracy of the estimates, and thus on the re-use. When we also dynamically adjust the transmitted power levels, bit allocations, and beam patterns we affect the interference seen by other users, and in effect worsen the channel dynamics, making estimation more difficult. However, these techniques provide interference avoidance and resistance to near/far effects. The question we (in part) address is how to balance these effects to produce a high performance design at moderate complexity.

III. Adaptive Multiple Access Techniques

In this section, we consider techniques such as channel probing, dynamic power, bit, and channel allocation, and adaptive antenna arrays as means to improve the capacity of slow-hopped TDMA systems. In our system, we presume that hopping patterns are arranged so that there is orthogonal channelization within cells, and randomization of patterns between cells. Thus, we are concerned with mitigating interference from outside the cell.

III.1 Dynamic Power and Channel Allocation (DPCA)

We now consider the joint optimization of power and channel assignments, beginning with the problem of optimal power allocation among users competing for one channel, here following [3]. Consider the situation depicted in Figure 1. Let G_{ij} be the propagation gain between the j^{th} base station and the mobile to which the i^{th} base station is communicating directly for some time slot. Let P_i be the power transmitted by the user belonging to base station i. If $P_i=0$, there is no mobile communicating with base station i in the slot. The SIR for basestation i is then given by

$$R_i = \frac{G_{ii}P_i}{\sum_{j \neq i} G_{ij}P_j} \qquad (1)$$

For reliable operation, we require $R_i > \gamma$, for all basestations. This set of conditions is equivalent to the following set of linear inequalities

$$G_{ii}P_i - \gamma \sum_{j \neq i} G_{ij}P_j > 0 \qquad (2)$$

where we must find whether there exists a set of positive powers $\{P_i\}$ for which all the inequalities are satisfied. It may be shown that this problem is equivalent to solving the matrix equation $\mathbf{G}\Pi=\mathbf{1}$, where Π is the vector of power assignments, $\mathbf{G}_{ii} = G_{ii}$, and $\mathbf{G}_{ij} = -\gamma G_{ij}$, with appropriate re-numbering of entries if not all basestations are active. Then if every component of the vector Π is positive, there is a feasible assignment. Moreover, the vector of powers is the best one to use apart from a scaling factor determined by the background noise, in the sense that it lies in the center of the region of feasible power assignments, and thus is least sensitive to perturbations.

Fully distributed forms of DPCA exist [4], in which the only information exchange required is that between a mobile and its own base station, revealing their measurements of the interference power and the propagation loss in that single link. If the SIR is below the target, the receiver signals the transmitter to increase the power by a small fixed amount in dB; if above the target, it signals for a decrease. This amounts to a gradient search on SIR, and has a geometric convergence rate. Specifically, all users adjust their power according to the following rule:

$$p_i(t+1) = \frac{p_i(t)\,\delta\gamma_i}{r_i(t)} \qquad (3)$$

where $p_i(t)$ is the power at time t, $r_i(t)$ is the measured SIR at time t, δ is a small safety margin to protect active users (those already admitted) and γ_i is the target SIR for user i. Numerous variations upon the rule in equation (3) are possible.

III.2 Channel Probing

One way to perform DPCA is to run the distributed algorithm for all the frequency hops, assigning power to the ones which can lead to an acceptable SIR. This is undesirable for two reasons: it disrupts the power levels for other users, and takes a long time, reducing robustness to mobility. Channel probing is a means to speed up the admission process while causing minimal disturbance to other users.

When the only additive impairment is noise, probing is simply a matter of learning the noise level and the channel impulse response. From these measurements, the best achievable SNR can be computed, and the appropriate combination of symbol rate, power level, and constellation size can be chosen. However, in systems with distributed adaptation of power, an increase in the power level of a new user will cause a reaction of increased power by the other users sharing the channel. The aim of probing is therefore not to measure the channel as it is, but to predict the SIR after the power control algorithm would have reached the steady state.

A very small number of measurements suffice to predict the final power level. Having measured the propagation gain (loss) on one's own link, a few power control steps establish the starting interference level and the change in interference as a result of increasing the power [5]. To probe the channel, let $q_i(0)$ be the measured received interference value before transmission begins. Let $q_i(1)$ be the measured interference value and $s_i(1)$ the measured received signal power value after a fixed power level has been transmitted, and the other users have adjusted their power in response. Then the maximum feasible SIR is estimated as

$$\hat{\gamma}_{max} = \frac{s_i(1)}{q_i(1) - q_i(0)} \qquad (4)$$

As there will be measurement errors, several iterations are needed in practice to form reasonable estimates of the final SIR, but probing is orders of magnitude faster than going through the complete adaptation. Since it may be accomplished at relatively low power levels, it results in minimal disruption of active users. Probing may also be used to predict the power level required to sustain any particular target SIR [5].

III.3 Dynamic Bit Allocation

It is not necessary to transmit the same number of bits in every hop, or indeed to use every hop. Consider for example a system in which hopping patterns have been randomized among cells. The digital constellation size is fixed, but we can choose in which M out of N slots in a hopping pattern to transmit bits. Coding is performed across the hops so that we are most concerned with the average power, although with a fixed constellation size minimum SIR requirements remain unless large redundancy is contemplated. Then the allocation of slots can be accomplished by using channel probing and choosing the M slots leading to the lowest power consumption. This choice causes users to arrange themselves for near-minimum mutual interference, i.e., it promotes interference avoidance [6]. Lower M permits more avoidance, but increases the bandwidth requirements for each user. Th optimal ratio of M/N can be determined by simulations.

Larger capacity is obtained if we can allocate a variable number of bits to the slots. The optimal allocation of bits and power when dealing with signal-independent impairments is the waterfilling distribution with respect to the noise floor . However, in multiple access systems the variable power allocation also changes the interference. In [7] it is shown that the allocation problem is NP hard, and so several ad hoc methods are attempted. The most successful are those which aim to achieve maximum interference avoidance, rather than a minimization of average transmitted power--that is, users tend to occupy as small a number of slots as possible consistent with the maximum constellation size permitted, with some smaller number using medium size constellations, and many with no bit allocation at all. Two algorithms which give good performance are reverse waterfilling based on estimates of SIR vs. transmitted power, and minimization of the total number of slots in the hopping pattern that are occupied, subject to having an acceptable SIR. In both cases, similar error rates are found in all hops for which bits are allocated, through the joint adjustment of constellation size and power level. Here accurate channel probing is essential, although channel coding over slots can permit reliable communications even if there remains some variability in uncoded error rates due to small errors in the SIR prediction.

III.4 Adaptive Antenna Arrays

Adaptive antenna arrays can provide a combination of diversity, directional gain, and interference nulling. This can be achieved using any number of iterative adaptive algorithms based on gradient descent or least squares, and comes at no expense in latency apart for the initial adaptation, in contrast to many other diversity and interference abatement techniques. However, a separate downconverter is required for each branch, which represents a large hardware and power consumption cost, and the antenna elements require some minimum spacing to provide independent channels. Thus, it would be desirable to increase the benefit of relatively small antenna arrays by using them both for transmission and reception.

Closed form solutions can be derived for the optimal weights of the receiver array given any particular transmitter pattern, but the joint optimization is already difficult for an adaptive transmitter array and receiver for a single link. The problem is even more difficult when the interferers also have adaptive arrays, since for strongly coupled links (large gains) an action by one produces a large reaction in the others. Figure 2 illustrates one simple scenario, where users (1,2) and (3,4) are the communications partners, and the barriers A-F are highly reflective. Some of the principal propagation paths are shown, with interference being caused only by multipath in this example. Suppose the link (1,2) is silent, and the users (3,4) have adaptive arrays but no equalizers. Then users 3 and 4 should adapt their arrays to suppress the multipath off barriers C and E and put high gain in the direct connection. However, suppose now that users 1 and 3 transmit in odd intervals and users 2 and 4 in even intervals. Then during reception user 3 must additionally suppress the interference from user 2, while user 1 need only be concerned by multipath from user 2 off barriers A and B. User 2 must suppress both multipath from its partner and multipath interference from user 3, while user 4 will be concerned only with multipath from user 3. If user 2 succeeds in nulling out the interference from user 3 and re-uses the same antenna pattern during transmission, then user 3 will experience very little interference from user 2 and can concentrate on multipath suppression. As there is no known closed form solution to the general form of this problem, the approach taken in [8] was adapting the receiver using the LMS algorithm, with receiver weights re-used for transmission. The heuristic behind this choice is that in a time division duplex system the nulls placed in the direction of interferers during reception will also reduce interference transmitted towards the other users who caused most of the interference. Thus, receiver adaptation will also serve to reduce the interference generated. Additionally, gain in the direction of the communications partner is preserved, so that SIR is improved for the link. These two effects also help to decouple interactions, promoting more rapid convergence.

One interesting result is that when users in the same cell are prohibited from using the same channel the system always converges. By restricting channel access in this way we insure that the dominant factor in adaptation is the desired link, rather than any particular interferer. Thus, such restrictions decouple the system, permitting rapid adaptation by least squares (gradient descent) methods. When this restriction is released, there are many occasions in which adaptation fails to converge. This is unsurprising, since there is no prior expectation that the error surface should be quadratic. One can also ensure stability by using an approach similar to that used in distributed power control algorithms, namely, voluntary dropout when little progress is being made in SIR levels. This mechanism also decouples the system, leading to stable convergence by the remaining users. Indeed, every distributed iterative method based on gradient descent will fail in some circumstances in multiple access systems, and so there must be some non-linear rule that either limits coupling in the first place, or makes a decision to halt adaptation by some subset of the users based on their priorities. That is to say, while in most circumstances gradient descent methods appear to work, we must either engineer the system so that the exceptions are extremely rare, and/or develop a procedure which determines when one of these exceptional circumstances is about to occur, so that a back-up procedure can be used (e.g., try another channel, go to a non-adaptive mode).

IV. Interaction of Adaptive Methods

We have so far described methods to adapt the channel assignment, power level, bit allocation, antenna beam pattern, with receiver-based or distributed algorithms, and probing to speed up the adaptation. Since there are conditions in which each technique will fail to provide the desired benefit, to design a robust system it is desirable to include a variety of these techniques. The question naturally arises to what extent they will work together or against each other, in a multiple access system coupled by mutual interference.

We consider a system in which the transmitter and receiver both have adaptive antenna arrays, adaptive equalization, and a wide dynamic range of power adjustment. Adjustment of either the power or the transmit beam pattern changes the interference seen by everyone else, and also modifies the impulse response seen by the intended receiver through selective weighting of multipath components. Adjustment of the receiver beam pattern changes the interference levels and the impulse response, and thus the power level required for reliable operation and the residual channel seen by the equalizer. The algorithms are therefore coupled at two levels: internally, with respect to the desired link, and externally, via the interference coupling to other users. There are many possible ways to combine power control, antenna arrays, and adaptive equalizers, not all of which are useful.

Adding equalizers to the system greatly expands the number of configurations possible. In multichannel combiners (MC) there are linear equalizers for each antenna branch, and then after combining, either MLSE or a DFE. Alternatively, all equalization can follow adaptive antenna combining. The MC deals with the distinct impulse responses seen in each antenna branch, and (in slowly changing channels) has better performance than structures where the equalizer follows simple combining. However, the complexity is higher and it is difficult to produce appropriate beam patterns for the transmitter array. For the indoor environment simulated in [8], there was little performance loss from using a simple combiner followed by equalization, so that the receiver antenna weights could be re-used for transmission.

Adding power control to the mix adds further interactions. A change in the beam pattern even with constant radiated power changes the power measured at all receivers. An interesting result is that if the antenna weights are renormalized to maintain constant gain in the direction of the desired users, then including small power control steps actually improves convergence of the distributed beamforming algorithm [1]. The intuition is that power control directly determines the minimum gain required for the desired link, quickly reducing unnecessary interference, and thus the coupling in adaptation among cells. We have earlier observed that enforcement of orthogonality of channel assignments within cells ensures the distributed antenna adaptation will converge. Such restrictions are not actually necessary so long as users will voluntarily terminate attempting to access a channel in the face of unsatisfactory improvements in the SIR. However, for the situation simulated in [1] capacity is actually roughly the same with or without the re-use restrictions, perhaps due to the high SIR required to support large constellations.

Probing to estimate the maximum SIR is made much more difficult by the presence of adaptive transmitter and receiver antenna arrays. When probing at low power levels, receivers can easily slightly shift their beam patterns to almost completely remove the effect of the new interference. Later, at some higher power level, shifting the beam pattern is insufficient and the transmitted power levels increase. Nevertheless, a simple model which assumes that most interference is due to one interference path predicts the final SIR with remarkable accuracy [9], at least for the situation of orthogonal channel assignments within cells. Thus, bit allocations are reasonably well predicted by probing, with relatively small adjustments to make after adaptation has been permitted to proceed.

V. Conclusion

The wireless multiple access channel is characterized by variability in space, time, frequency, and interference. The design of an efficient communications system is largely an exercise in the management of this variability. It is our ability to estimate the state of the channel (including interference) which most fundamentally limits network capacity. The channel and interference can be engineered to a considerable extent through resource re-use restrictions, separation of users according to locations or traffic types, and the use of adaptive transmission and reception techniques. These latter techniques exploit variations in the radio channel with angle, frequency, and location, but to yield higher capacity must necessarily deal with much stronger interference coupling in the network than conventional designs with fixed re-use patterns. Much of the time distributed algorithms based on gradient descent techniques suffice to produce stable convergence, but mechanisms are required to recognize conditions which may lead to instability, and then to deal with these situations using a back-up procedure. In effect, we apply one set of rules to the main part of the probability distribution, and another set to the (pathological) tails, with some effort spent in recognizing the situation (e.g., through probing). While at this time many of the solutions are ad hoc, in aggregate they promise large capacity improvement using distributed iterative algorithms.

Our motivation in producing the above set of algorithms was to provide a set of compatible techniques for low-complexity implementation. However, the interference interactions in and of themselves are very interesting. There appears to be room for a more rigorous mathematical analysis from the point of view of coupled dynamical systems, the results of which may suggest new algorithmic approaches to stable and rapid convergence. Additionally, for bursty traffic techniques which rely upon receiver/transmitter cooperation are ineffective due to the large overhead relative to message size, and the rapid interference dynamics. Thus, the optimal multiple access scheme for this type of traffic is radically different from the system outlined above. How to most effectively deal with mixed traffic in multiple access wireless systems remains a wide open topic.

Acknowledgements

The results reported here are based upon the research of my students Christopher Hansen, Eldad Perahia, and Charles Wang. Thanks also go to Desmond Taylor for arranging a much needed sabbatical at the University of Canterbury to spend some time contemplating the coupled adaptation problem.

References

[1] Eldad Perahia. Diversity Combining, Adaptive Antennas, and Equalization for Digital Radio. Ph.D. Dissertation, Electrical Engineering Dept., University of California, Los Angeles, 1995.

[2] H.L. Van Trees. Detection, Estimation, and Modulation Theory. Wiley: New York, 1968.

[3] N. Bambos and G.J. Pottie, “Power control based admission policies in cellular radio networks”, IEEE Globecom, Dec. 6-9, 1992, Orlando FL, pp. 863-867.

[4] S. Chen, N. Bambos, and G.J. Pottie, “Radio link admission algorithms for wireless networks with power control and active link protection,” technical report UCLA-ENG-94-25

[5] C.J. Hansen, C.C. Wang, and G.J. Pottie, “Distributed dynamic channel resource allocation in wireless communication systems,” 28th. Asilomar Conf. on Signals, Systems, and Computers, Oct. 1994 pp. 78-82.

[6] C.C. Wang and G.J. Pottie, “Interference avoidance and power control strategies for coded frequency hopped cellular systems,” Proc. ICC ‘95.

[7] C.C. Wang and G.J. Pottie, "Bit algorithms for FH-CDMA wireless communication systems," 34th Allerton Conf. on Commun., Control, and Computing, 1996.

[8] E. Perahia and G.J. Pottie, “Adaptive antenna arrays and equalization for indoor digital radio,” Proc. ICC ‘96.

[9] Christopher J. Hansen. Probing Techniques for Multiuser Channels with Power Control. Ph.D. Dissertation, Electrical Engineering Dept., University of California, Los Angeles, 1997.

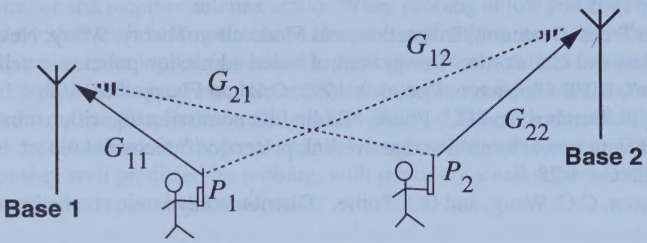

G21
G12
G11
G22
P1
P2
Base 1
Base 2

Figure 1: Propagation scenario between mobiles and base stations.

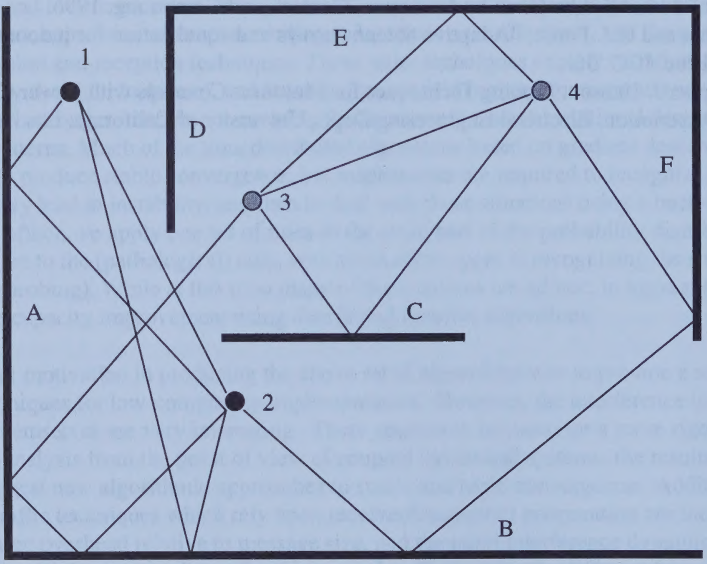

1
2
3
4
A
B
C
D
E
F

Figure 2: The interference-coupled beamforming problem

Blind Adaptive Space-Time Linear Multiuser Demodulation

Michael L. Honig[†], Rapeepat Ratasuk[†], and Sergio Verdú[‡]

[†]Department of ECE, Northwestern University
2145 Sheridan Road, Evanston, IL 60208

[‡]Department of ECE, Princeton University
Princeton, New Jersey 08544

Abstract

We consider blind demodulation of multiple digital signals in a cellular environment under the assumption that training data is unavailable, and that the receiver has no *a priori* knowledge of user signatures in space or time. Our approach to this problem is to exploit diversity in the time, frequency, and spatial domains through the application of a linear receiver with multiple antenna elements. To separate the users, a subspace decomposition is used to obtain an initial estimate of the users' signatures. Once estimates of the signatures are available, the anchored algorithm described in [1] can be used to obtain initial bit estimates. These bit estimates are then used as the input to a decision-directed least squares (LS) algorithm (either full-rank or reduced-rank), which provides refined bit estimates for the packet of interest. This process can be iterated until the bit estimates remain fixed. Simulation results indicate that for the Time-Division Multiple-Access (TDMA) model considered, low to moderate uncoded error rates are achievable when the number of antennas is sufficiently large (at least twice the number of users), and when users differ substantially in received power (at least 6 dB).

1 Introduction

We consider blind demodulation of multiple digital signals in a cellular environment. The receiver is assumed to be randomly located within a particular cell, and is unable to cooperate with the transmitters. The emphasis of our study is on *adaptive multiuser* demodulation with limited side information. Although some training information may be available to aid the demodulation (e.g., the midamble in GSM), here we assume that no training data is available, that the receiver has no prior knowledge of signatures (in time or space), and that near-far effects are likely due to the random location of the receiver relative to the transmitters.

The proposed algorithm assumes a linear receiver with multiple antenna elements. Our approach is to first obtain an initial estimate of the spatial signatures via a subspace decomposition. This type of signal space decomposition has been previously considered for channel estimation [2], interference suppression [3], [4], and timing estimation [5], [6].

Once estimates of the spatial signatures are available, the anchored algorithm described in [1] can be used to obtain initial bit estimates. These bit estimates are then used as the input to a decision-directed least squares (LS) algorithm proposed in [7] , which provides refined bit estimates for the packet of interest. A related blind multiuser demodulation algorithm which uses maximum-likelihood estimation is given in [8]. This process can be iterated until the bit estimates remain fixed. By increasing the number of dimensions available through diversity, we simplify the blind multiuser demodulation problem, resulting in algorithms which have good performance with moderate complexity. In addition, diversity provides robustness in the presence of low Signal-to-Noise Ratios and fading.

In the next section, we briefly describe the TDMA model under consideration. In Section 3, we present a brief description of linear multiuser demodulation, followed by a detailed discussion of the blind algorithm in Section 4. In Section 5, we discuss enhancements to the algorithm as well as its extension to multipath channels. Simulation results are then presented in Section 6. Finally, we summarize our findings in Section 7.

2 System Model

The communications scenario considered assumes TDMA, so that the multiple users being demodulated are in different cells. We remark that, in the scenario considered here, the number of users to be demodulated is quite small (i.e., less than five), and the primary source of diversity is in the space domain. The received signal corresponding to the k^{th} user is

$$s_k(t) = \sum_i b_k(i) g(t - iT - \tau_k) \tag{1}$$

where $b_k(i)$ is the i^{th} transmitted symbol for user k, $g(t)$ is the transmitted pulse shape, T is the symbol period, and τ_k is the random delay associated with the k^{th} user. To alleviate phase slips associated with Rayleigh fading, we assume that $b_k(i)$ is differently encoded. The received signal at the m^{th} antenna is then

$$r_m(t) = \sum_m \sum_k h_k^{(m)}(t) \otimes s_k(t) + n_m(t) \tag{2}$$

where $h_k(t)$ is the channel impulse response from user k to antenna m, $n_m(t)$ is the noise on antenna m, and "$\otimes$" denotes convolution. In the model considered, we assume that each antenna element fades independently. In the case of flat fading, the spatial signature for a particular user is the vector of complex fade coefficients across the antennas. The signatures are assumed to be constant throughout each packet, but are assumed to be statistically independent from packet to packet.

For each antenna at the receiver we assume a filter matched to $g(t)$ followed by a sampler. With M antennas, the vector of M output samples at time iT can be written as

$$\mathbf{r}(i) = \mathbf{H}b(i) + \mathbf{n}(i) \tag{3}$$

where the mth component corresponds to antenna m, and the (m, k)th component of the $M \times K$ matrix $\mathbf{H}$ is the fading coefficient associated with the path from user k to antenna m (K being the number of users).

3 Linear Multiuser Demodulation

Given the sequence of received vectors, $\{\mathbf{r}(i)\}$, our objective is to obtain an estimate of the source bits transmitted by all users of interest. In what follows, we assume that we know *a priori* the number of users to be demodulated. Since a linear receiver is assumed, the decision statistic for user k at time i is given by

$$z_k(i) = \mathbf{C}_k^\dagger \mathbf{r}(i) \tag{4}$$

where $\mathbf{C}_k$ represents the filter coefficients associated with user k. Assuming DPSK modulation, the estimate of the source symbol is then

$$\hat{b}_k(i) = \text{sign}\{\text{Re}(z_k^*(i-1)z_k(i))\} \tag{5}$$

The performance criterion used to optimize the receiver is Mean Squared Error (MSE), defined as

$$\epsilon_k = E(|b_k(i) - \mathbf{C}_k^\dagger \mathbf{r}(i)|^2) \tag{6}$$

The filter which minimizes the MSE (Minimum MSE (MMSE) solution) is

$$\mathbf{C} = \mathbf{R}^{-1}\mathbf{P} \tag{7}$$

where $\mathbf{C}$ is the matrix with column vectors $\mathbf{C}_k$, $k = 1, \ldots, K$,

$$\mathbf{R} = E[\mathbf{r}(i)\mathbf{r}^\dagger(i)] = \mathbf{H}\mathbf{H}^\dagger + \sigma_n^2 \mathbf{I}, \tag{8}$$

σ_n^2 is the noise variance at each sampling instant, and the k^{th} column of $\mathbf{P}$ is the signature of user k given by

$$\mathbf{P}_k = E[b_k(i)\mathbf{r}(i)] = \mathbf{H}_k \tag{9}$$

4 Blind Demodulation Algorithm

In the absence of a training sequence, blind techniques such as the constant modulus algorithm can be used; however, these tend to converge slowly [9]. Our approach is to obtain an initial estimate of the signatures via a subspace decomposition, and subsequently to substitute these estimates into an LS solution for $\mathbf{C}$. This algorithm can converge quite rapidly, although the performance will depend on the accuracy of the initial signature estimates.

4.1 Initial Signature Estimates

If the signatures corresponding to the different users are orthogonal (i.e., if $\rho_{kl} = \mathbf{H}_k^\dagger \mathbf{H}_l = 0$, $k \neq l$), then the signatures are eigenvectors of the covariance matrix $\mathbf{R}$. Although this is unlikely to be true when the elements of the signatures are independent random variables, the larger the dimension of $\mathbf{H}_l$, the lower the expected cross-correlation ρ_{kl} will be. An initial estimate of the signatures can therefore be obtained by computing the eigenvectors of the estimate of $\mathbf{R}$,

$$\widehat{\mathbf{R}} = \frac{1}{B}\sum_{i=1}^{B} \mathbf{r}(i)\mathbf{r}^\dagger(i) \tag{10}$$

where B is the number of symbols per time slot.

Since $\widehat{\mathbf{R}}$ is positive definite and full rank, we can write

$$\widehat{\mathbf{R}} = \mathbf{V}\Lambda\mathbf{V}^{\dagger} \tag{11}$$

where $\mathbf{V}$ is the orthogonal matrix with columns given by the eigenvectors of $\mathbf{R}$, and Λ is the diagonal matrix of eigenvalues arranged in descending order

$$\lambda_1 > \lambda_2 > \dots \lambda_K > \sigma^2, \quad \lambda_n = \sigma^2, \quad n > K \tag{12}$$

Since we assume that the number of users to be demodulated (K) is known, the signatures are initially taken to be the K eigenvectors $\mathbf{V}_1, \dots, \mathbf{V}_K$ corresponding to the largest eigenvalues.

4.2 Reduced-Rank LS Algorithm

Given the signature estimates, we form a reduced column matrix $\widetilde{\mathbf{V}} = [\mathbf{V}_1, \dots, \mathbf{V}_K]$ and project the received signal onto the corresponding signal subspace to obtain

$$\widetilde{\mathbf{r}}(i) = \widetilde{\mathbf{V}}'\mathbf{r}(i) \tag{13}$$

(In what follows, all variables with a tilde are reduced-rank, or projected, variables.) We wish to select $\widetilde{\mathbf{C}}$ to minimize

$$\epsilon = E|\mathbf{b}(i) - \widetilde{\mathbf{C}}'\widetilde{\mathbf{r}}(i)|^2 \tag{14}$$

The solution is given by

$$\widetilde{\mathbf{C}} = \widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{P}} \tag{15}$$

where

$$\widetilde{\mathbf{R}} = E[\widetilde{\mathbf{r}}(i)\widetilde{\mathbf{r}}^{\dagger}(i)] = \widetilde{\mathbf{V}}'\mathbf{R}\widetilde{\mathbf{V}} = \widetilde{\Lambda} \tag{16}$$

and the projected signature for user k is

$$\widetilde{\mathbf{P}}_k = E[b_k(i)\widetilde{\mathbf{V}}'\mathbf{r}(i)] = \widetilde{\mathbf{V}}'E[b_k(i)\mathbf{r}(i)] = \widetilde{\mathbf{V}}'\mathbf{H}_k \tag{17}$$

Hence, we can rewrite the blind reduced-rank filter as

$$\widetilde{\mathbf{C}} = \widetilde{\Lambda}^{-1}\widetilde{\mathbf{V}}'\mathbf{H}_k, \qquad \widetilde{\Lambda} = \mathit{diag}[\lambda_1, \dots, \lambda_K] \tag{18}$$

If the signatures are orthogonal, then $\widetilde{\mathbf{V}}'\mathbf{H}_k = ||\mathbf{V}_k||^2\mathbf{e}_k$, where $\mathbf{e}_k$ is the kth unit vector. An LS algorithm is obtained by replacing the expectations in (14)-(17) by time averages.

Our initial estimates of the users' signature assume that they are orthogonal, whereas they are only approximately orthogonal. Hence, subsequent to obtaining the initial bit estimates $\{\widehat{b}_k(i)\}$, we can refine the estimate of the signatures by estimating the cross-correlation in 17. With differential decoding, the bit decisions obtained from (5) can be used to obtain the signature estimate as follows [7]:

$$\widehat{\widetilde{\mathbf{P}}}_k = \frac{1}{B}\sum_{i=1}^{B}[\widehat{b}_k(i)z_k(i-1)]^*\widetilde{\mathbf{r}}(i) \tag{19}$$

This in turn gives a new set of filter coefficients $\widetilde{\mathbf{C}} = \widetilde{\Lambda}^{-1}\widehat{\widetilde{\mathbf{P}}}_k$. Note that any phase offset in $z_k(i)$ is introduced into the signature estimate, but this is inconsequential with differential decoding. Using the new set of filter coefficients, we compute new estimates of the transmitted bits. This procedure is iterated until the algorithm converges. (That is, the estimated bits do not change from one iteration to the next.) The algorithm is not guaranteed to converge, and problems with convergence do occur at relatively high error rates (> 15%). However, simulations indicate that the iterative procedure does improve the error rate when it is less than 10%.

4.3 Full-Rank LS Algorithm

The full-rank LS algorithm also relies on estimates of the user signatures via the initial eigen-decomposition. However, the estimated bits are used to minimize the "full-rank" cost function

$$\epsilon_k = \sum_{i=1}^{B} w^{-(B-i)} |b_k(i) - \mathbf{C}_k^{\dagger}\mathbf{r}(i)|^2 \tag{20}$$

where w is an exponential weighting constant. The $\mathbf{C}$ which minimizes the cost function, modified for differential detection, is $\mathbf{C} = \widehat{\mathbf{R}}^{-1}\widehat{\mathbf{P}}$ where

$$\widehat{\mathbf{R}} = \sum_{i=1}^{B} w^{-(B-i)}\mathbf{r}(i)\mathbf{r}^{\dagger}(i), \quad \widehat{\mathbf{P}}_k = \sum_{i=1}^{B} w^{-(B-i)}\hat{b}_k(i) z_k^*(i-1)\mathbf{r}(i) \tag{21}$$

The initial bit estimates are then used to recalculate $\mathbf{C}$ according to the LS criterion, resulting in new bit estimates. The process is iterated until the algorithm converges. For the examples considered, simulation results indicate that the full-rank algorithm performs better than the reduced-rank algorithm when the initial error rate is ≤ 10%.

5 Algorithmic Enhancements

In this section, we present enhancements to the basic algorithm described in the preceding section. We also discuss the effect of multipath, and propose a modification of the differentially coherent algorithm which noncoherently combines estimates from adjacent symbols.

5.1 Fractional Sampling

In addition to spatial oversampling, we can also oversample the received signal in time. However, in contrast to spatial oversampling, the performance gain due to oversampling in time is limited by the signal bandwidth. For the numerical results that follow, we assume that $g(t)$ is a raised-cosine pulse with 50% excess bandwidth. In that case, the Nyquist sampling rate is less than two samples per symbol.

It is convenient to represent the received samples during a symbol interval as the matrix $\mathbf{r}(i) = [\mathbf{r}(i;1)\ \mathbf{r}(i;2)]$ where $\mathbf{r}(i;n)$ denotes the n^{th} sample of the i^{th} sampling interval. The use of multiple samples per bit provides robustness against timing offset which is a concern in asynchronous communications system. Although fractional sampling introduces correlation into the noise samples, numerical results indicate that this causes only a slight degradation in performance.

5.2 Multi-symbol Observation Window

The discussion in the preceding section assumed that the filter spanned a single symbol interval in time. Of course, it is conceptually straightforward to expand this window to span multiple symbols in time. This is especially important when ISI and multipath are present. For the numerical results in the next section, we assume that the receiver is asynchronous. Since $g(t)$ is a raised-cosine pulse shape, all users experience ISI even without multipath. Expanding the observation window allows the filter to capture the additional energy due to multipath.

5.3 Soft Decision-Directed Adaptation

From the discussion in Section 4, the signature corresponding to user k can be estimated as

$$\widehat{\mathbf{P}}_k = \frac{1}{B}\sum_{i=1}^{B} \mathcal{F}(\mathbf{C}_k^\dagger \mathbf{r}(i))\mathbf{r}(i) \tag{22}$$

where $\mathcal{F}(x) = \mathrm{sgn}(x)$ in the case of hard decisions and coherent detection. To improve performance, we can weight the decisions by a measure of reliability. One approach is to select $\mathcal{F}(x)$ to

$$\min_{\mathcal{F}} E(|\mathcal{F}(\mathbf{C}_k^\dagger \mathbf{r}(i)) - \mathbf{C}_k^\dagger \mathbf{r}(i)|^2) \tag{23}$$

where it is assumed that

$$\mathbf{C}_k^\dagger \mathbf{r}(i) = b_k(i) + \zeta_k(i) \tag{24}$$

where $\zeta_k(i)$ is a Gaussian random variable with mean zero and variance

$$\sigma_k^2 = \mathrm{MSE} \approx \frac{1}{B}\sum_{i=1}^{B} |\hat{b}_k(i) - \mathbf{C}_k^\dagger \mathbf{r}(i)|^2 \tag{25}$$

The solution to this optimization problem, presented in [10], is

$$\mathcal{F}(x) = \tanh(\frac{x}{\sigma_k^2}) \tag{26}$$

The hyperbolic tangent function discounts samples that are close to zero while providing more weight to those that are close to ±1. The numerical results in Section 6 were generated with soft decisions (modified for differential detection), although in most cases the performance improvement due to using soft decisions, instead of hard decisions, is relatively minor.

5.4 Multipath

To illustrate the effect of multipath, consider a two-ray channel in which the rays are separated by the symbol interval T. The received signal is then described by

$$\mathbf{r}(i) = \mathbf{H}^1\mathbf{b}(i) + \mathbf{H}^2\mathbf{b}(i-1) + \mathbf{n}(i) \tag{27}$$

and the signature of user k is comprised of $\mathbf{H}_k^1$ corresponding to the main path, and $\mathbf{H}_k^2$ corresponding to the second path.

The MMSE criterion automatically combines the two paths coherently. By combining received vectors from two symbols, we are presented with a new received vector

$$\begin{bmatrix} \mathbf{r}(i) \\ \mathbf{r}(i+1) \end{bmatrix} = \begin{bmatrix} \mathbf{H}^2 & \mathbf{H}^1 & 0 \\ 0 & \mathbf{H}^2 & \mathbf{H}^1 \end{bmatrix} \begin{bmatrix} \mathbf{b}(i-1) \\ \mathbf{b}(i) \\ \mathbf{b}(i+1) \end{bmatrix} + \begin{bmatrix} \mathbf{n}(i) \\ \mathbf{n}(i+1) \end{bmatrix} \tag{28}$$

The signature of user k is now given by $\mathbf{P}_k = [\mathbf{H}_k^1\ \mathbf{H}_k^2]'$ and a signal subspace decomposition can be used as before to separate the users. Specifically, compute the new autocorrelation matrix and decompose it according to

$$\widehat{\mathbf{R}} = \frac{1}{B} \sum_{i=1}^{B} \begin{bmatrix} \mathbf{r}(i) \\ \mathbf{r}(i+1) \end{bmatrix} \begin{bmatrix} \mathbf{r}(i) \\ \mathbf{r}(i+1) \end{bmatrix}^{\dagger} = \mathbf{V}\Lambda\mathbf{V}^{\dagger} \tag{29}$$

The user signatures are then given by

$$\widehat{\mathbf{H}}_k = \mathbf{V}_k, k = 1, \ldots, K \tag{30}$$

Note that the intersymbol interference due to multipath is equivalent to additional multiple access interference.

If the filter spans only a single symbol interval, then noncoherent equal gain combining (e.g., see [11]) can be used to combine the paths associated with adjacent symbol intervals. Specifically, the algorithm relies on the assumption that the different paths are received with independent fading coefficients on each receiver antenna. Thus, each interfering path can be viewed as an independent interferer. If the different paths are sufficiently strong, then a signal subspace (eigen) decomposition should be able to resolve the paths. For example, with two users, each with two paths, the eigendecomposition will provide four distinct eigenvectors that span the signal subspace. The eigenvalues associated with a particular user are associated with the received path energies. Of course, the multipath components must be strong relative to the background noise level. Otherwise, the associated eigenvalues are too small to allow the multipath subspace to be distinguished from the noise subspace.

Subsequent to path resolution, equal gain noncoherent combining can be used to estimate the bits. With this method, the estimated signatures associated with each path for a particular user are used to compute soft differential estimates of the data, which are added together to obtain the decision statistic. This technique is analogous to a noncoherent RAKE receiver for Direct-Sequence (DS)-Code- Division Multiple Access (CDMA) with the main difference being that the delayed time-domain CDMA signatures associated with the different paths are replaced by space-domain signatures.

6 Numerical Results

The results in this section assume a fixed number of users (two or three), with the users at different (random) locations. The modulation scheme is binary DPSK with raised cosine pulses with 50% excess bandwidth. The length of the time slot is 150 bits, and the receiver is asynchronous with respect to each user, with each user delayed by an independent random variable. Therefore $\mathbf{r}(i)$ contains both intersymbol and multiple-access interference. Both frequency-nonselective and frequency-selective (two-ray) Rayleigh fading are considered. The fading coefficients associated with the paths and antennas are assumed to be independent, and are constant for the duration of the time slot. Where fractional sampling is used, the noise correlation is taken into account.

6.1 Frequency-Nonselective Fading Channels

Unless otherwise noted, the results presented here assume that the signal-to-noise ratio for user one is 10 dB. For two users, the power difference between the strong and weak user is 8 dB. The MMSE solution sometimes gave zero errors over the duration of the simulation, in which case the results are not plotted in the figures.

Figure 1 shows results for the MMSE, blind reduced-rank, and blind full-rank algorithms assuming two samples per bit. For both users the blind algorithms perform significantly worse than the ideal MMSE solution. The full-rank LS algorithm performs slightly better than the reduced-rank LS algorithm in this case. A comparison of results obtained from sampling once and twice per symbol is shown in Figure 2. The performance improvement associated with fractional sampling is due to robustness with respect to random timing offset.

In Figure 3, error rate is shown as a function of the difference in power between the users. When the powers of the two users are close, the blind algorithm is unable to separate the users since the eigenvectors of the sample covariance matrix are no longer aligned with each user's signature.

Finally, Figure 4 illustrates the performance of the algorithm when 3 users are present. The SNR for user 1 is 11 dB. The powers of users 2 and 3 are 8 dB and 11 dB below that of user 1, respectively. When fractional sampling is used, the performance of user 1 is close to that shown in the preceding figures. The performance of the weak users, however, are substantially worse. This is due to the fact that the two weak users are close together in power. Hence, the subspace decomposition is unable to produce reliable estimates of the weak signatures.

6.2 Frequency-Selective Channels

Figures 5-6 show the performance of the algorithm in the presence of frequency-selective fading. We consider a two-path model with the second path at equal strength and delayed by the symbol interval T. Figure 5 illustrates the performance based on coherent combining of multipath signals. In this case the MMSE results are better than the analogous results for frequency-nonselective fading performance shown in Figure 1, since the MMSE solution coherently combines the different paths. The performance of the blind algorithms degrades significantly when multipath is present. This is because the additional ISI from multipath degrades the initial signature estimates.

The results for noncoherent combining using a window of two symbols is presented in Figure 6. These results show that noncoherent combining performs slightly better than coherent combining.

7 Conclusions

Our numerical results indicate that for the cases considered, the blind algorithm performs well when the number of antennas is sufficiently large (at least twice the number of users), and when users differ substantially in received power (at least 6 dB). The use of fractional sampling is important mainly due to the insensitivity provided to timing offset. We also remark that the error rates shown here are for the uncoded symbols. Coded error rates would be substantially lower, so that intelligible reception of the weak user may be possible (assuming voice transmission) when the uncoded error rate is quite high (e.g., 10%).

Perhaps the main weakness of the blind algorithm proposed is that it is unable to separate the users when the received powers are close. In addition, we have assumed that the number of users to be demodulated is known *a priori.* If this is not the case, then one possibility is to demodulate users in decreasing order of received power until the weakest user demodulated is unintelligible. This, of course, assumes that the number of active users is not too large.

Finally, our preliminary results indicate that severe multipath is likely to cause substantial degradation in performance, especially for relatively weak users. Enhancements to the algorithm which may improve performance in this situation are currently being studied.

References

[1] M. L. Honig, U. Madhow, and S. Verdu, "Blind Adaptive Multiuser Detection," *IEEE Transactions on Information Theory*, vol. 41, pp. 944–960, July 1995.

[2] H. Liu and G. Xu, "Smart Antennas in Wireless Systems: Uplink Multiuser Blind Channel and Sequence Detection," *IEEE Transactions on Communications*, vol. 45, pp. 187–199, Feb 1997.

[3] X. Wang and H. V. Poor, "Blind Multiuser Detection: A Subspace Approach," *preprint:1997.*

[4] M. L. Honig, "A Comparison of Subspace Adaptive Filtering Techniques for DS-CDMA Interference Suppression," in *Proc. 1997 Milcom*, (Monterey, Ca), 1997.

[5] S. E. Bensley and B. Aazhang, "Subspace-Based Channel Estimation for CDMA Communication Systems," *IEEE Transactions on Communications*, vol. 44, pp. 1009–1020, Aug. 1996.

[6] E. G. Strom, S. Parkvall, S. L. Miller, and B. E. Ottersten, "Propagation Delay Estimation of DS-CDMA Signals in a Fading Environment," *Proc. 1994 IEEE Globecom/Comm. Theory Mini-Conf.*, pp. 85–89, Dec. 1994.

[7] M. L. Honig, M. J. Shensa, S. Miller, and L. Milstein, "Performance of Adaptive Linear Interference Suppression for DS-CDMA in the Presence of Flat Rayleigh Fading," *IEEE/VTC 47th Vehicular Technology Conference*, pp. 2191–2195, May 1997.

[8] S. Talwar, M. Viberg, and A. Paulraj, "Blind Separation of Synchronous Co-Channel Digital Signals Using an Antenna Array-Part I: Algorithms," *IEEE Transactions on Signal Processing*, vol. 44, pp. 1184–1197, May 1996.

[9] J. J. Treichler and M. G. Larimore, "New Processing Techniques Based on the Constant Modulus Adaptive Algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 420–431, Apr. 1985.

[10] R. Vijayan and H. V. Poor, "Nonlinear Techniques for Interference Suppression in Spread-Spectrum Systems," *IEEE Transactions on Communications*, vol. 38, pp. 1060–1065, July 1990.

[11] H. Huang and S. Verdú, "Linear Differentially Coherent Multiuser Detection for Multipath Channels," *Wireless Personal Communications*, 1997.

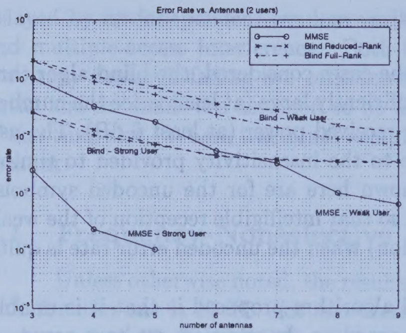


Figure 1: Error rate vs. number of antennas, 2 users, 2 samples/bit, $\text{SNR}_1 = 10$ dB, $\text{SNR}_2 = 2$ dB.

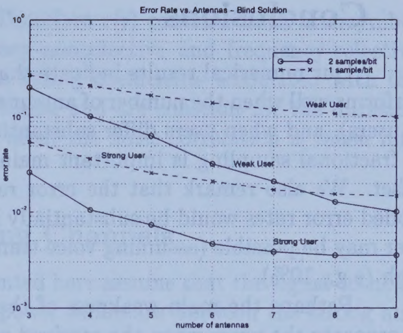


Figure 2: Fractional Sampling: Blind reduced-rank solution, 2 users, $\text{SNR}_1 = 10$ dB, $\text{SNR}_2 = 2$ dB.

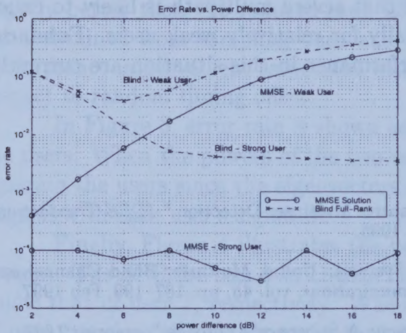


Figure 3: Error rate vs. relative power, 2 users, 5 antennas, 2 samples/bit, $\text{SNR}_1 = 10$ dB, $\text{SNR}_2 = 2$ dB.

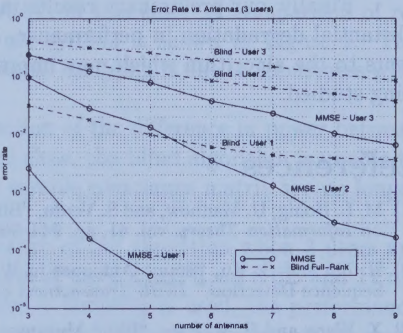


Figure 4: Error rate vs. number of antennas, 3 users, 2 samples/bit, $\text{SNR}_1 = 11$ dB, $\text{SNR}_2 = 3$ dB, $\text{SNR}_3 = 0$ dB.

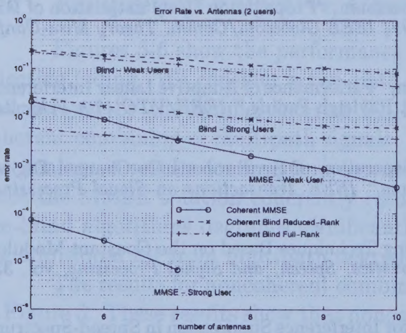


Figure 5: Coherent Combining: 2 samples/bit, $\text{SNR}_1 = 10$ dB, $\text{SNR}_2 = 2$ dB.

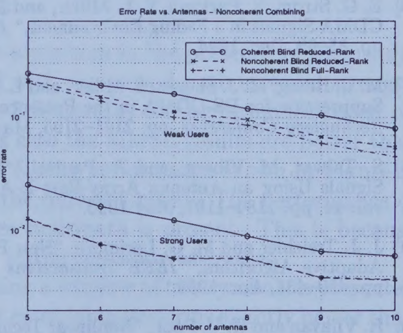


Figure 6: Noncoherent Combining: window = 2 symbols, 2 samples/bit, $\text{SNR}_1 = 10$ dB, $\text{SNR}_2 = 2$ dB.

Joint Transmitter and Receiver Beamforming for Maximum Capacity in Spatial Division Multiaccess

J.-H. Chang, L. Tassiulas, and F. Rashid-Farrokhi

Electrical Engineering Department and Institute for Systems Research
University of Maryland
College Park, MD 20742, USA

Abstract

Antenna array beamforming and power control have been studied separately to increase the capacity in wireless communication networks. Recently, the two were considered jointly in a system with power control capabilities which uses antenna array at the receiver only or the transmitter only. In this paper we formulate and study the joint beamforming and power control problem in a system with antenna arrays both at the transmitter and the receiver. Two algorithms are proposed for the joint problem. The capacity of the system that has beamforming capabilities at both sides is compared with the system that has one side beamforming capabilities only. The increase in the network capacity is evaluated through a numerical study.

1 Introduction

The antenna diversity combining has been used as a means to increase the capacity in wireless communication networks[1]. Power control has been studied to minimize the transmitter power and reduce the interference to the other users[2],[3]. Power control and diversity combining problem has been recently considered jointly with beamforming at base stations only; either the receivers[4] or the transmitters[5]. A distributed algorithm that converges to the optimal power vector and beamforming vectors was proposed for the joint power control and receiver beamforming, and a distributed algorithm that converges to a feasible solution was proposed for the joint power control and transmitter beamforming.

In this paper, we study the capacity increase obtained by considering antenna array beamforming and power control jointly. The joint beamforming and power control problem is formulated for the case when both the transmitter and the receiver have antenna arrays. Two algorithms are proposed for the joint problem, and the capacity of the system that has beamforming capabilities at both sides is compared with the system that has one side beamforming capabilities only. The increase in the network capacity is evaluated through a numerical study.

The paper is organized as follows: The system model is described and the joint beamforming and power control problems are formulated in section 2. For each problem defined, two corresponding algorithms are proposed in section 3. Simulation results are presented in section 4. Finally, conclusion is in section 5.

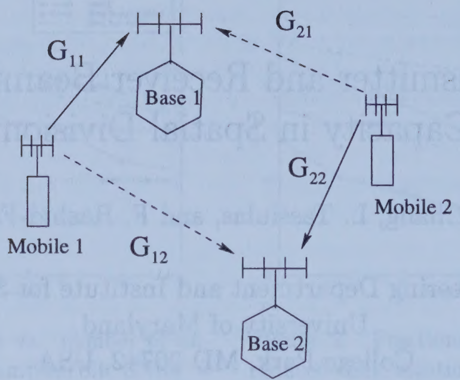


Figure 1: System model showing two pairs of transmitters and receivers

2 Transmitter-Receiver Beamforming and Power Control

Suppose there are M transmitters communicating with M corresponding receivers. Each receiver has a linear adaptive antenna array with K_R elements, and each transmitter has a linear adaptive antenna array with K_T elements. Figure 1 depicts two pairs of mobile transmitters and base station receivers with 4 antenna array elements each, and the path gain between mobile i and the base j is denoted as G_{ij}. To help understanding, the uplink case will be considered throughout the paper. The uplink power vector is denoted as $\mathbf{P}$ where P_i is the ith mobile's transmitter power, for $i = 1, 2, ..., M$. The receiver antenna array weight vector at base i is $\mathbf{w}_i$, and the transmitter antenna array weight vector at mobile i is $\hat{\mathbf{w}}_i$. The array response vector of the receiver antenna array at base i for mobile j is $\mathbf{a}_{ji}$, and the array response vector of the transmitter antenna array at mobile j for the base i is $\tilde{\mathbf{a}}_{ij}$. The array response of base j for the mobile i can be represented as a matrix of dimension $K_R \times K_T$,

$$\mathbf{A}_{ji} = \mathbf{a}_{ji}\tilde{\mathbf{a}}_{ij}^T. \tag{1}$$

Figure 2 shows jth transmitter antenna array and ith receiver antenna array. The array response matrix is referred to as $\mathbf{A}_{ji}$ where $\mathbf{A}_{ji}[k][l]$ denotes the array response between the lth array element of the mobile j and kth element of the base i.

The array element outputs are linearly combined with weights to give the received signal. The received signal at the ith base station receiver is

$$x_i = \sum_{k=1}^{K_R} w_i^*[k] r_i[k] \tag{2}$$

where w^* denotes the complex conjugate of w, and $r_i[k]$ is the received signal at the kth element of ith receiver antenna array given by

$$r_i[k] = \sum_{j=1}^{M} \sum_{l=1}^{K_T} A_{ji}[k][l] \hat{w}_j^*[l] \sqrt{P_j G_{ji}} s_j + n_i[k] \tag{3}$$

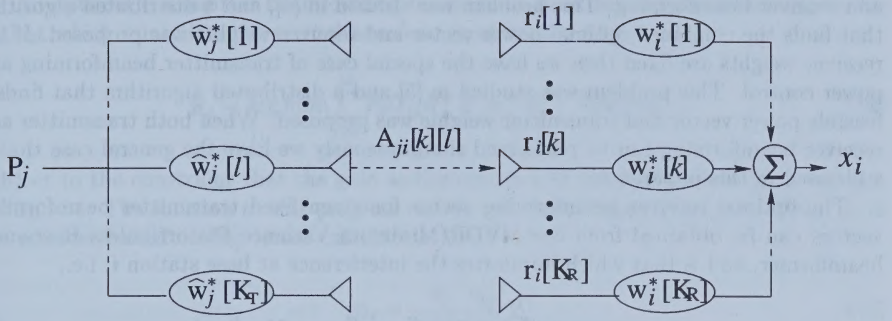


Figure 2: Antenna arrays of jth transmitter and ith receiver and the array response matrix $\mathbf{A}_{ji}$

where s_j is the message signal transmitted from mobile j and $n_i[k]$ is the thermal noise at the input of the kth array element of receiver i. Inserting (3) into (2) gives

$$\begin{aligned} x_i &= \textstyle\sum_{k=1}^{K_R} w_i^*[k](\sum_{j=1}^{M}\sum_{l=1}^{K_T} A_{ji}[k][l]\hat{w}_j^*[l]\sqrt{P_jG_{ji}}s_j + n_i[k]) \\ &= \textstyle\sum_{j=1}^{M} \mathbf{w}_i^H \mathbf{A}_{ji}\hat{\mathbf{w}}_j^*\sqrt{P_jG_{ji}}s_j + \mathbf{w}_i^H\mathbf{n}_i \end{aligned} \tag{4}$$

where $\mathbf{w}_i, \hat{\mathbf{w}}_j$ are the beamforming weight vectors at the ith receiver and the jth transmitter respectively, $\mathbf{n}_i$ is the thermal noise vector at the ith receiver, and $\mathbf{w}^H$ denotes Hermitian of the matrix $\mathbf{w}$. Assuming that the noise and the signals are uncorrelated and zero-mean, the average received signal plus interference power is

$$\begin{aligned} \mathbf{E} &= E[x_i x_i^H] \\ &= \textstyle\sum_{j=1}^{M} \mathbf{w}_i^H\mathbf{A}_{ji}\hat{\mathbf{w}}_j^*\{\mathbf{w}_i^H\mathbf{A}_{ji}\hat{\mathbf{w}}_j^*\}^H P_jG_{ji} + \mathbf{w}_i^H E[\mathbf{n}_i\mathbf{n}_i^H]\mathbf{w}_i \\ &= \textstyle\sum_{j=1}^{M} \mathbf{w}_i^H\mathbf{A}_{ji}\hat{\mathbf{w}}_j^*\hat{\mathbf{w}}_j^T\mathbf{A}_{ji}^H\mathbf{w}_iP_jG_{ji} + N_i\mathbf{w}_i^H\mathbf{w}_i \end{aligned} \tag{5}$$

where N_i is the noise power at the input of each array element. Therefore, the SIR at the ith base station receiver is given by

$$\Gamma_i = \frac{P_iG_{ii}\mathbf{w}_i^H\alpha_{ii}\alpha_{ii}^H\mathbf{w}_i}{\sum_{j\neq i}^{M} P_jG_{ji}\mathbf{w}_i^H\alpha_{ji}\alpha_{ji}^H\mathbf{w}_i + N_i\mathbf{w}_i^H\mathbf{w}_i} \tag{6}$$

where α_{ji} is defined as $\alpha_{ji} = \mathbf{A}_{ji}\hat{\mathbf{w}}_j^*$.

The following two problems are considered related to the model above.

Problem 1 *Maximize the minimum SIR over all receivers, for all combinations of power vector* $\mathbf{P}$, *receiver beamforming vector* $\mathbf{w}_i$, *and transmitter beamforming vector* $\hat{\mathbf{w}}_i$, *for all* i.

Problem 2 *Minimize the sum of transmitter powers* $\sum_i P_i$, *subject to the SIR constraints,* $\Gamma_i \geq \gamma_0$, *for all* i, *where* γ_0 *is an achievable "system protection ratio".*

If the transmitter weights are fixed then we have the special case of joint power control and receiver beamforming. This problem was studied in [4], and a distributed algorithm that finds the combined optimal power vector and receiver weights was proposed. If the receiver weights are fixed then we have the special case of transmitter beamforming and power control. This problem was studied in [5] and a distributed algorithm that finds a feasible power vector and transmitter weights was proposed. When both transmitter and receiver beamforming can be performed simultaneously we have the general case that is addressed in this paper.

The optimal receiver beamforming vector for given fixed transmitter beamforming vectors can be obtained from the MVDR(Minimum Variance Distortionless Response) beamformer, and is that which minimizes the interference at base station i, i.e.,

$$\underline{\mathbf{w}}_i = \arg\min_{\mathbf{w}_i}\Big(\sum_{j\neq i}^{M} P_j G_{ji}\mathbf{w}_i^H \alpha_{ji}\alpha_{ji}^H \mathbf{w}_i + N_i \mathbf{w}_i^H \mathbf{w}_i\Big), \tag{7}$$

subject to the constraint that the gain at the direction of interest is equal to unity, i.e., $\mathbf{w}_i^H \alpha_{ii} = 1$. The solution to this problem is given by[7]

$$\underline{\mathbf{w}}_i = \frac{\Phi_i^{-1}\alpha_{ii}}{\alpha_{ii}^H \Phi_i^{-1}\alpha_{ii}} \tag{8}$$

where Φ_i is the correlation matrix of the received vector $\mathbf{r}_i$. If the message signals s_j are orthonormal to each other, the correlation matrix Φ_i is given by[6]

$$\Phi_i = \sum_{j=1}^{M} P_j G_{ji}\alpha_{ji}\alpha_{ji}^H + N_i\mathbf{I}. \tag{9}$$

For fixed transmitter and receiver beamforming vectors, define $\mathbf{D}$ and $\mathbf{F}$ as

$$\mathbf{D} = diag\{\frac{1}{G_{11}}, ..., \frac{1}{G_{MM}}\}, \tag{10}$$

and

$$\mathbf{F}_{ij} = \begin{cases} G_{ji}\|\mathbf{w}_i^H \mathbf{A}_{ji}\hat{\mathbf{w}}_j^*\|^2 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{11}$$

When $\mathbf{D}$ and $\mathbf{F}$ are fixed the problem is the same as the power control with omnidirectional antennas. The maximum of the minimum SIR over all receivers can be calculated by

$$\gamma_{max} = \frac{1}{\rho(\mathbf{DF})} \tag{12}$$

where $\rho()$ denotes the Perron-Frobenius eigenvalue, and the minimum power vector is the eigenvector of the corresponding eigenvalue, i.e.,

$$\mathbf{P} = \gamma_{max}\mathbf{DFP}. \tag{13}$$

For transmitter beamforming, the *virtual reverse link* concept is used[5] where the transmitters and receivers interchange locations. Let $\tilde{P}_i$ denote the virtual downlink power of base station i. The virtual downlink signal to interference ratio at mobile i is given by

$$\tilde{\Gamma}_i = \frac{\tilde{P}_i G_{ii}\hat{\mathbf{w}}_i^H \tilde{\alpha}_{ii}\tilde{\alpha}_{ii}^H \hat{\mathbf{w}}_i}{\sum_{j\neq i}^{M} \tilde{P}_j G_{ij}\hat{\mathbf{w}}_i^H \tilde{\alpha}_{ji}\tilde{\alpha}_{ji}^H \hat{\mathbf{w}}_i + \tilde{N}_i \hat{\mathbf{w}}_i^H \hat{\mathbf{w}}_i} \tag{14}$$

where $\tilde{\alpha}_{ji}$ is defined as $\tilde{\alpha}_{ji} = \mathbf{A}_{ij}^T \mathbf{w}_j^*$, and $\tilde{N}_i$ is virtual noise power at transmitter i.

An optimal transmitter beamforming vector for a fixed receiver beamforming vector is that which minimizes the virtual interference at mobile i, i.e.,

$$\underline{\hat{\mathbf{w}}}_i = \arg\min_{\hat{\mathbf{w}}_i} \Big(\sum_{j \neq i}^{M} \tilde{P}_j G_{ij} \hat{\mathbf{w}}_i^H \tilde{\alpha}_{ji} \tilde{\alpha}_{ji}^H \hat{\mathbf{w}}_i + \tilde{N}_i \hat{\mathbf{w}}_i^H \hat{\mathbf{w}}_i\Big), \tag{15}$$

subject to the constraint that the gain at the direction of interest is equal to unity, i.e., $\hat{\mathbf{w}}^H \tilde{\alpha}_{ii} = 1$. The solution to this problem can be calculated using a similar method as in the receiver beamforming vector calculation, i.e.,

$$\underline{\hat{\mathbf{w}}}_i = \frac{\tilde{\Phi}_i^{-1} \tilde{\alpha}_{ii}}{\tilde{\alpha}_{ii}^H \tilde{\Phi}_i^{-1} \tilde{\alpha}_{ii}} \tag{16}$$

where $\tilde{\Phi}_i$ can be called a virtual correlation matrix and is given by

$$\tilde{\Phi}_i = \sum_{j=1}^{M} \tilde{P}_j G_{ij} \tilde{\alpha}_{ji} \tilde{\alpha}_{ji}^H + \tilde{N}_i \mathbf{I}. \tag{17}$$

3 Proposed Algorithms

In this section, two algorithms are proposed for problems 1 and 2 respectively.

3.1 The Algorithm for Problem 1

Algorithm 1

1. Calculate the maximum achievable SIR for fixed receiver and transmitter weights and compute $\mathbf{P}$ that maximizes the minimum SIR over all receivers, i.e.,

$$\gamma_{max} = \frac{1}{\rho(\mathbf{DF})}$$

and

$$\mathbf{P} = \gamma_{max} \mathbf{DFP}.$$

2. Calculate receiver weights to maximize SIR for fixed powers and transmitter weights

$$\underline{\mathbf{w}}_i = \arg\min_{\mathbf{w}_i} \Big(\sum_{j \neq i}^{M} P_j G_{ji} \mathbf{w}_i^H \alpha_{ji} \alpha_{ji}^H \mathbf{w}_i\Big),$$

subject to $\mathbf{w}_i \alpha_{ii} = 1$, for $i = 1, 2, ..., M$.

3. Calculate the maximum achievable SIR for fixed receiver and transmitter weights and compute $\tilde{\mathbf{P}}$ that maximizes the minimum SIR over all virtual downlink transmitters, i.e.,

$$\gamma_{max} = \frac{1}{\rho(\mathbf{DF^T})}$$

and

$$\tilde{\mathbf{P}} = \gamma_{max} \mathbf{DF}^T \tilde{\mathbf{P}}.$$

4. Calculate transmitter weights to maximize SIR for fixed virtual downlink powers and receiver weights

$$\underline{\hat{\mathbf{w}}}_i = \arg\min_{\hat{\mathbf{w}}_i}\Big(\sum_{j\neq i}^{M} \tilde{P}_j G_{ij} \hat{\mathbf{w}}_i^H \tilde{\alpha}_{ji}\tilde{\alpha}_{ji}^H \hat{\mathbf{w}}_i\Big),$$

subject to $\hat{\mathbf{w}}_i\tilde{\alpha}_{ii} = 1$, for $i = 1, 2, ...M$.

5. Go to step 1 and repeat the iteration.

It can be shown that algorithm 1 always converges to a power vector and corresponding receiver/transmitter vectors. As a result a corresponding system protection ratio γ_{max} is achieved. This value, γ_{max} can be used to compare the capacity of the network having beamforming capability at one side only with the network having beamforming capability at both sides of the transmitter and the receiver. The larger γ_{max} is the larger the capacity of the network is.

3.2 The Algorithm for Problem 2

A distributed and iterative algorithm for problem 2 is proposed. The algorithm can be viewed as a combination of the algorithms in [4] and [5]. The power update iteration is motivated by the fact that if there exists a feasible power vector for a fixed set of transmitter and receiver weights, then the iteration converges.

Algorithm 2

1. Calculate receiver weights to maximize SIR for fixed powers and transmitter weights

$$\underline{\mathbf{w}}_i = \arg\min_{\mathbf{w}_i}\Big(\sum_{j\neq i}^{M} P_j G_{ji} \mathbf{w}_i^H \alpha_{ji}\alpha_{ji}^H \mathbf{w}_i + N_i \mathbf{w}_i^H \mathbf{w}_i\Big)$$

subject to $\mathbf{w}_i^H \alpha_{ii} = 1$.

2. Update power

$$P_i^{n+1} = \gamma_0 \sum_{j\neq i} \frac{G_{ji}\underline{\mathbf{w}}_i^H \alpha_{ji}\alpha_{ji}^H \underline{\mathbf{w}}_i}{G_{ii}} P_j^n + \frac{\gamma_0 N_i \underline{\mathbf{w}}_i^H \underline{\mathbf{w}}_i}{G_{ii}}$$

3. Calculate transmitter weights to maximize SIR for fixed powers and receiver weights

$$\underline{\hat{\mathbf{w}}}_i = \arg\min_{\hat{\mathbf{w}}_i}\Big(\sum_{j\neq i}^{M} \tilde{P}_j G_{ij} \hat{\mathbf{w}}_i^H \tilde{\alpha}_{ji}\tilde{\alpha}_{ji}^H \hat{\mathbf{w}}_i + \tilde{N}_i \hat{\mathbf{w}}_i^H \hat{\mathbf{w}}_i\Big)$$

subject to $\hat{\mathbf{w}}_i^H \tilde{\alpha}_{ii} = 1$.

4. Update virtual reverse link power

$$\tilde{P}_i^{n+1} = \gamma_0 \sum_{j\neq i} \frac{G_{ij}\underline{\hat{\mathbf{w}}}_i^H \tilde{\alpha}_{ji}\tilde{\alpha}_{ji}^H \underline{\hat{\mathbf{w}}}_i}{G_{ii}} \tilde{P}_j^n + \frac{\gamma_0 \tilde{N}_i \underline{\hat{\mathbf{w}}}_i^H \underline{\hat{\mathbf{w}}}_i}{G_{ii}}$$

5. Go to step 1 and repeat the iteration.

As the result of this algorithm a set of power vector and receiver weights and transmitter weights are found. The total transmitted power,

$$\sum_{i=1}^{M} P_i \hat{\mathbf{w}}_i^H \hat{\mathbf{w}}_i \tag{18}$$

can be calculated. Note that if γ_0 is not achievable then algorithm 2 diverges.

4 Simulation Results

A numerical study of algorithms 1 and 2 was performed. A CDMA network is considered where M users are randomly distributed in the area $[0.5, 3.5] \times [0.5, 3.5]$ with uniform distribution and 9 base stations are located at integer grids. The link gain is modeled as $G_{ij} = 1/r^4$, where r is the distance between the mobile and the base station. The shadow fading is ignored in these simulations. The transmitter and the receiver antenna arrays have 4 elements each.

To simulate the system with omnidirectional antennas only, we can run the algorithms while fixing all the antenna array weights at both the transmitters and the receivers at $[1, 0, ..., 0]^T$. For receiver beamforming only system, all the transmitter array weights can be fixed at $[1, 0, ..., 0]^T$.

Figure 3 shows the achievable SIR, γ_{max} versus the total number of users in the network. For the system with only receiver antenna array and the system with omnidirectional antennas the achievable SIR is indeed the maximum achievable SIR. The system with both transmitter and receiver beamforming has higher achievable system protection ratio than the other two cases. For a given SIR constraint γ_0, the system with beamforming at both sides can support more users than the receiver beamforming only system. For example, for $\gamma_0 = 0.1$ receiver beamforming only system can support up to 230 users while the system with both sides beamforming capability can support up to 300 users.

Algorithm 2 was simulated with the same network settings with the SIR constraint $\gamma_0 = 0.0304$. Figure 4 shows the total transmitted power versus the number of users in the system. While the system with no beamforming capability and the system with the beamforming capability at the receiver only could support 200 and 800 users respectively, the system with both transmitter and receiver beamforming capabilities could support up to 1000 users. One can also observe the savings in the total transmitted power by a factor of 4 or more.

5 Conclusion

Joint power control and diversity combining was considered in case where both transmitters and receivers have antenna arrays. Two related problems were formulated and two corresponding algorithms were proposed and the capacity increase obtained from the transmitter and receiver beamforming and power control was evaluated. With antenna arrays both at the transmitter and the receiver, larger system protection ratio was achievable than the other cases, which means larger capacity. For a given feasible signal to noise ratio constraint, savings in total transmitted power was also observed.

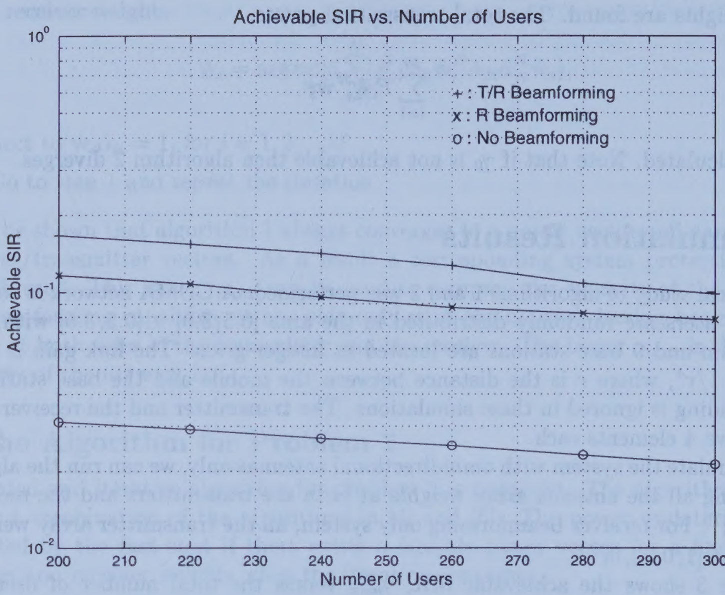


Figure 3: A comparison of achievable SIR vs. number of users

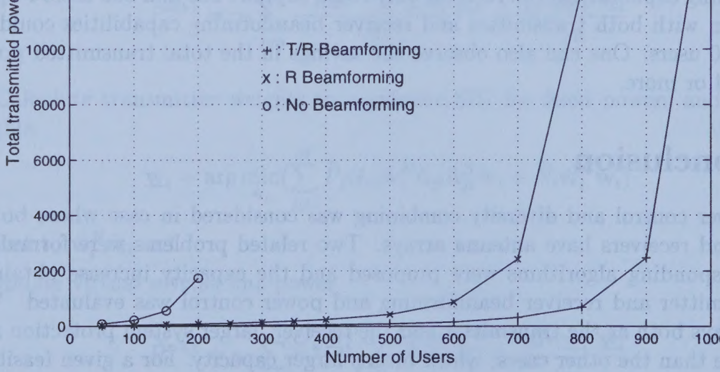


Figure 4: Total transmitted power vs. number of users

References

[1] A.F. Naguib, A. Paulraj, and T. Kailath, "Capacity Improvement with Base-station Antenna Arrays in Cellular CDMA," *IEEE Trans. Vehicular Technol.*", vol. 43, No. 3, Aug. 1994.

[2] S. Grandhi, R. Vijayan, D.J. Goodman, and J. Zander, "Centralized power control for cellular radio systems," *IEEE Trans. Vehicular Tech.*, vol.42, Nov. 1993.

[3] J. Zander, "Distributed Cochannel Interference Control in Cellular Radio Systems," *IEEE Trans. Vehicular Technol.*, vol. 41, Aug. 1992.

[4] F. Rashid-Farrokhi, L. Tassiulas, K.J.R. Liu, "Joint Optimal Power Control and Beamforming in Wireless Networks Using Antenna Arrays," *ISR Technical Report*, Institute for Systems Research, University of Maryland at College Park, Aug. 1996.

[5] F. Rashid-Farrokhi, L. Tassiulas, K.J.R. Liu, "Transmit Beamforming for Cellular Wireless Communications," *Conference on Information Sciences and Systems (CISS-97)*, March 1997.

[6] B. Suard, A. Naguib, G. Xu, and A. Paulraj, "Performance analysis of CDMA mobile communication systems using antenna array," *Proc. ICASSP'93*, vol. IV, Minneapolis, MN, pp.153-156, April 1993.

[7] R.A. Monzingo and T.W. Miller, *Introduction to Adaptive Arrays.* New York Wiley, 1980.

QUANTIZATION FOR DISTRIBUTED ESTIMATION WITH COMMUNICATION AND STORAGE CONSTRAINTS

Vasileios Megalooikonomou and Yaacov Yesha
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250

Abstract

We consider the problem of quantizer design in a distributed estimation system subject not only to communication constraints at the channels but also to storage constraints at the fusion center in the case where the observation statistics are unknown and only a training sequence is available. Our scheme involves growing and pruning of regression trees along with some labeling techniques for iteratively decreasing the estimation error. We propose two approaches in order to reduce the storage requirements at the fusion center. The first approach gives a direct sum estimate of the parameter. The second approach is based on neural networks and gives more flexibility in terms of the number of parameters used. Simulations show the trade-offs between storage complexity of the fusion center and performance of the quantizers in the two cases.

1 Introduction

A common scheme of a distributed estimation system consists of a single fusion center where an estimation of some quantity has to be made and of remote sensors that observe the environment and transmit relevant information to the fusion center. The capacity constraints on the communication lines along with the storage constraints on the fusion center suggest some very challenging problems.

The goal is to minimize the expected parameter estimation error. Due to the capacity constraints at the communication lines, the data from the sensors are transmitted at rates insufficient to convey all the observations reliably. Therefore, the observations are vector quantized and the estimation is achieved via compressed information.

The sensors cannot communicate with each other and there is no feedback from the fusion center back to them. The communication channels are assumed to be error free. We assume fixed length coding for the transmission. Although the number of sensors can be in general arbitrary here we consider the two-sensor case since the same solution can be easily extended to the more general case. The observations at the sensors are random variables. Here, we consider the case where the joint probability density function is unknown.

A solution for the problem of quantizer design for a distributed estimation system in the case of unknown observation statistics has been given by Megalooikonomou and

Yesha [7]. The same problem in the case of known probability model was considered by Lam and Reibman [5, 6]. However, storage complexity issues for the fusion center were not addressed in those papers. Both methods use a simple table for the fusion center that is indexed by the codewords of the two quantizers. Storage requirements of this fusion center table can be a problem for a large number of sensors and/or a large number of codewords for each sensor. Gubner [3] considers the problem of quantizer design for this system subject not only to communication constraints but also to computation constraints at the fusion center. Gubner assumes that the probability model is known. The computational capabilities of the fusion center are constrained to direct sum estimation of the parameter. As a result, the algorithm uses only bivariate joint distributions. Since, in our case, the distribution is unknown, Gubner's method cannot be used.

Here, we consider the problem of quantizer design subject to both communication constraints and storage constraints at the fusion center in the case where the joint probability model is unknown and one must rely on a training set. We present two approaches that use reduced storage representations of the fusion center table. The first approach gives a direct sum estimation of the parameter. The second approach uses a neural network for the representation of the fusion center table. It gives more flexibility in terms of the number of parameters used to represent the fusion center and on the achievable performance than the first approach.

2 Background

Let X_1^q and X_2^r be the random observation vectors at the sensors and θ the unobservable continuous quantity that the fusion center tries to estimate. We use the vector notation X_k^p, for the sensor k, as a shorthand for $(X_k[1], X_k[2], \ldots, X_k[p])$. Let $\{(X_1^q, X_2^r)^{(t)}, \theta^{(t)}; t = 1, \ldots, M\}$ be the training set, $\mathcal{T}$ of size M that represents the statistics of the source, Q_k be the quantizer for sensor k, and $\hat{X}_k^{p,t}$ be the codeword for the observation $X_k^{p,t}$ that is transmitted to the fusion center. The task of the fusion center is to estimate the unobserved quantity θ based on the $\hat{X}_k^{p,t}$ it receives. Let h be the function of the fusion center that gives the estimate of θ and $P_{Q_1} = \{U_i; i = 1, \ldots, N\}$ and $P_{Q_2} = \{V_j; j = 1, \ldots, L\}$, be the partition regions for the quantizers Q_1 and Q_2, respectively. The transmitted values, $\hat{X}_1^{q,t}$ and $\hat{X}_2^{r,t}$ from the sensors to the fusion center are given by:

$$\hat{X}_1^{q,t} = \sum_{i=1}^{N} (i-1) I_{U_i}(X_1^{q,t}) \tag{1}$$

$$\hat{X}_2^{r,t} = \sum_{j=1}^{L} (j-1) I_{V_j}(X_2^{r,t}) \tag{2}$$

where $I_A(x)$ denotes the indicator function of a set $A \subset \Re^d$ of dimension d, i.e. $I_A(x) = 1$ if x is in A and $I_A(x) = 0$ otherwise. Then the output of the decentralized estimation system is $h(\hat{X}_1^{q,t}, \hat{X}_2^{r,t})$.

Let also $\mathcal{R}_i$, $\mathcal{R}'_j$, and $\mathcal{R}_{i,j}$ be the following subsets of the training set:

$$\mathcal{R}_i = \{(X_1^q, X_2^r)^{(t)} : X_1^{q,t} \in U_i\} \tag{3}$$

$$\mathcal{R}'_j = \{(X_1^q, X_2^r)^{(t)} : X_2^{r,t} \in V_j\} \tag{4}$$

$$\mathcal{R}_{i,j} = \{(X_1^q, X_2^r)^{(t)} : X_1^{q,t} \in U_i, X_2^{r,t} \in V_j\} \tag{5}$$

Then the fusion center h has the following value for each pair of codewords i, j:

$$h(i,j) = \frac{1}{|\mathcal{R}_{i,j}|} \sum_{t:(X_1^q, X_2^r)^{(t)} \in \mathcal{R}_{i,j}} \theta^{(t)} \tag{6}$$

We consider the mean-squared error (mse) distortion function. The objective is to find Q_1, Q_2, and h such that the error expression below is minimized

$$Error = \frac{1}{M} \sum_{t=1}^{M} \left(\theta^{(t)} - h(\hat{X}_1^{q,t}, \hat{X}_2^{r,t}) \right)^2 \tag{7}$$

We design the quantizers using the regression tree [1] approach proposed by Megalooikonomou and Yesha [7] in order to deal with the problem of unknown joint distribution. Their design is cooperative and involves some growing, pruning, and labeling techniques. They build a regression tree for each one of the sensors and then reduce the rate using a labeling technique.

The regression trees are decision trees with queries of the form $X_k[i] < c_j$ (for an observation variable $X_k[i]$ and a constant c_j) where each leaf is labeled by an estimation value which is generally constant. All splits are on single variables so they are perpendicular to the coordinate axes. For observations of dimension d the leaves of the regression tree correspond to d-dimensional rectangles (intervals in the 1-dimensional case). The constraint that is imposed when building the regression trees is that one tree cannot have different splits based on answers to queries on the other tree. This is because the encoders do not communicate with each other and they do not have both X_1^q and X_2^r.

2.1 Growing and pruning of the regression trees

The growing of the regression trees is done as follows. The tree for X_1^q (X_2^r) is grown given no information about X_2^r (X_1^q) and then the tree for X_2^r (X_1^q) is grown taking into account the already built tree for X_1^q (X_2^r). The growing of the tree is based on the decrease of the error in the estimation of the parameter θ. Pruning is also involved in the growing procedure because in order to grow right size trees, instead of attempting to stop the splitting at the right set of terminal nodes someone has to continue the splitting until the expectation of the estimation error is below a certain threshold (resulting in a large tree) and then selectively prune (by recombining leaves that are siblings) this large tree.

The basic operation that finds the best split point for one tree given the other tree considers the existing rectangles of the other tree. This is reflected in the calculation of the estimation error that is performed using Equation 7. For sensor k we choose the next split (considering all the variables $X_k[1], \ldots, X_k[p]$ and all the values of these variables) that minimizes the estimation error given by Equation 7 that depends on the existing rectangles of both trees. Also, when one tree is pruned given the other tree, the risk of every node (i.e. the expected estimation error of the subtree that has as root that node [1]) in the tree is calculated taking into account the other tree.

When both trees, T_1, T_2, have been built and their pruned subtrees, S_1, S_2 have been found, they can be improved by first growing from the beginning the tree for X_1^q getting a new tree T_1' given the pruned tree S_2, then pruning the tree T_1' getting the tree S_1' and then performing the following:

1. Continue pruning both trees, starting always from the original trees, T_1' and T_2. The pruning for the tree T_2 can proceed in a different direction given the different tree S_1' for X_1^q. The same holds for the pruning of the tree T_1'.

2. Instead of pruning the original trees, grow from the beginning each tree and prune it back given the pruned tree for the other sensor.

After growing the trees, the rectangles are labeled using an algorithm that is related to the Cyclic Generalized Lloyd Algorithm (CGLA), the s-CGLA, in order to combine the rectangles into the required number of partition regions. Then the trees are grown from the beginning in order to improve over the previous trees including in this procedure the labels that have been assigned to the rectangles. In order to do that one tree is kept fixed along with its labels and the other tree is grown given this tree. The fusion center table h and the estimation error in this case is calculated using Equations 6 and 7 where the rectangles of the first tree that have the same label are grouped into the same partition region. Some initial labels are assigned to the rectangles and these labels are iteratively improved using again the s-CGLA, keeping during the whole phase the other tree including its labels fixed. The same procedure is repeated for the other tree. Both trees are repeatedly improved until the improvement in the estimation error is below a certain threshold.

The pruning algorithm that is used is the *recursive optimal pruning algorithm* (ROPA) proposed by Kiang *et al* [4]. Their algorithm is an extension of the Generalized BFOS (GBFOS) algorithm proposed by Chou *et al* [2] that optimally prunes a TSVQ codebook and which is a generalization of an algorithm by Breiman *et al* [1] for classification and regression trees. The purpose of the pruning of the original regression trees in the case of fixed rate quantization is to get a subtree with a given number of leaves and with estimation error that is as small as possible.

2.2 Labeling of the rectangles

Labeling is used to denote the assignment of codewords to the rectangles of the regression trees in the sense that rectangles that have the same label form a quantizer partition region and use the same codeword for the transmission. Megalooikonomou and Yesha [7] introduce two algorithms related to the CGLA, the s-CGLA and the lh-s-CGLA. Both algorithms consider together groups of training samples. The second algorithm changes the fusion center temporarily whenever there is a decision that has to be made in order to calculate the effect of every possible change. Moreover, the second algorithm keeps the fusion center table updated all the time. The resulting quantizers can have either connected or disconnected partition regions.

Let n_k be the number of codewords and $m_k \geq n_k$ be the number of leaves for quantizer k. Let also $l(r)$ be the label of a specific rectangle r. Given the partition regions P_{Q_1} and P_{Q_2}, for X_1^q and X_2^r respectively, the optimal fusion center h is given by:

$$h(m,n) = \frac{1}{|\mathcal{R}'_{m,n}|} \sum_{t:\left(X_1^q, X_2^r\right)^{(t)} \in \mathcal{R}'_{m,n}} \theta^{(t)} \tag{8}$$

where $\mathcal{R}'_{m,n}$ is the following subset of the training set:

$$\mathcal{R}'_{m,n} = \{(X_1^q, X_2^r)^{(t)} : l(r(X_1^{q,t})) = m,\ l(r(X_2^{r,t})) = n\} \tag{9}$$

The estimation error, err_i, contributed by the following subset, $\mathcal{R}''_i$, of the training set (based on X_1^q)

$$\mathcal{R}''_i = \{(X_1^q, X_2^r)^{(t)} : r(X_1^{q,t}) = i\} \tag{10}$$

is given by:

$$err_i = \frac{1}{|\mathcal{R}''_i|} \sum_{t:(X_1^q, X_2^r)^{(t)} \in \mathcal{R}''_i} \left(\theta^{(t)} - h(l(i), l(r(X_2^{r,t})))\right)^2 \tag{11}$$

The total estimation error is then given by:

$$Error = \sum_{i:0\ldots m_1-1} err_i \tag{12}$$

The estimation error can also be expresed using a similar formula and the corresponding subsets based on X_2^r.

The main component of lh-s-CGLA performs the following for each sensor k [7] until the reduction on the estimation error given by Equation 12 is less than a given threshold:

```
for (each rectangle $i$ from 0 to $m_k - 1$) {
    for (each label $j$ from 0 to $n_k - 1$) {
        $l(i) \leftarrow j$;
        calculate $h$ using Equation 8;
        calculate estimation error, $err_i[j]$, using Equation 11;
    }
    $l(i) \leftarrow \arg\min_{j:0\ldots(n_k-1)}(err_i[j])$;
    calculate $h$ using Equation 8;
}
```

The breakpoint initialization method is used to initialize the labels of the rectangles. This method initializes the labels by first pruning even more the final pruned subtrees, F_1 and F_2 to the number of labels n_1 and n_2 getting the trees, F_1' and F_2' and then assigning the same label to all the rectangles of the trees F_1 and F_2 that correspond to one rectangle of the trees F_1' and F_2'.

3 Direct sum estimation of the parameter

Constraining the storage requirements of the fusion center and recalling that $\hat{X}_1^q = i - 1$ if and only if X_1^q is in U_i and $\hat{X}_2^r = j - 1$ if and only if X_2^r is in V_j we require (as in [3]) that the estimation of the parameter θ for a certain training point t is

$$\hat{\theta}^{(t)} = \sum_{i=1}^{N} a_i I_{U_i}(X_1^{q,t}) + \sum_{j=1}^{L} b_j I_{V_j}(X_2^{r,t}) \tag{13}$$

where $a_i : i = 1, \ldots, N$ and $b_j : j = 1, \ldots, L$ are the parameters of the fusion center that we are trying to find. If the partition regions of the sensors are fixed, choosing the a_i's, b_j's that minimize

$$Error_{lin} = \frac{1}{M} \sum_{t=1}^{M} \left(\theta^{(t)} - \hat{\theta}^{(t)}\right)^2 \tag{14}$$

is a linear estimation problem whose solution is given by the normal equations. We use an iterative solution to the normal equations.

The parameters $a_i, i : 1, \ldots, N$ are given as follows:

$$a_i = \frac{1}{|\mathcal{R}_i|}\left(\sum_{j=1}^{L} \sum_{t:\left(X_1^q, X_2^r\right)^{(t)} \in \mathcal{R}_{i,j}} \theta^{(t)} - \sum_{j=1}^{L} |\mathcal{R}_{i,j}| b_j\right) \tag{15}$$

A similar formula is derived for the parameters $b_j, j : 1, \ldots, L$:

$$b_j = \frac{1}{|\mathcal{R}'_j|}\left(\sum_{i=1}^{N} \sum_{t:\left(X_1^q, X_2^r\right)^{(t)} \in \mathcal{R}_{i,j}} \theta^{(t)} - \sum_{i=1}^{N} |\mathcal{R}_{i,j}| a_i\right) \tag{16}$$

We calculate the parameters $a_i, i : 1, \ldots, N$ and $b_j, j : 1, \ldots, L$ iteratively using Equations 15, 16, until the reduction on the estimation error given by Equation 14 is less than a given threshold. This iterative method solves the normal equations using a training set and converges to the actual solutions given by analytical methods. The method that we use to solve the normal equations is actually the Gauss-Seidel (successive displacements) method where each new component is immediately used in the calculation of the next component.

When the partition regions of one sensor along with the fusion center parameters are fixed we partition the space for the other quantizer using a variation of the methods for growing and pruning the regression trees that were used in the case of an unrestricted fusion center [7]. This variation of the methods is imposed by the restrictions on the storage requirements of the fusion center. The estimation error is now calculated using Equation 14. Also, when we build the regression tree for X_1^q taking into account the partitions of the quantizer for X_2^r along with their associated parameters $b_j, j : 1, \ldots, L$ we use the quantities

$$\theta^{(t)} - b_j$$

where $X_2^{r,t}$ is in partition region V_j ($I_{V_j}(X_2^{r,t})$), instead of $\theta^{(t)}$ in the methods that try to find the best split on X_1^q. A similar variation is used when we build the regression tree for X_2^r given the regression tree for X_1^q. It can be proved that minimizing the expected squared estimation error based on the above quantities is equivalent to minimizing the expected squared estimation error based on the direct sum estimate. The proof is easy and is omitted. A crucial point is that in the case of direct sum estimation, finding the best splits for one tree, when the other tree and the coefficients that correspond to it are fixed, is feasible. If we do not take into account the restrictions on the fusion center when building the regression trees and we only apply the procedure that calculates the parameters of the fusion center as the last step of our scheme the performance is worse. In other words, it is better to build the quantizers taking into account also the special form that the fusion center function takes in order to satisfy the storage constraints.

4 A neural network representation of the fusion center

We can further improve the performance achieved by the direct sum estimation of the parameter using a neural network representation of the fusion center table (i.e. of the reconstructed values given by Equation 6).

The neural network that we use is a two-layer feed-forward network and the learning rule is backpropagation with momentum and adaptive learning rate. The momentum method decreases the probability that the network will get stuck in a shallow minimum in the error surface and helps decrease training times. Adaptive learning rate decreases training time by keeping the learning rate reasonably high while insuring stability. For the first layer we use a hyperbolic tangent transfer function and for the second layer we use a linear transfer function. This kind of networks has been proven capable of approximating any function with a finite number of discontinuities with arbitrary accuracy. By varying the number of neurons, i.e. the number of weights and biases (number of parameters) of the neural network, we achieve various trade-offs of the complexity of the fusion center and the performance of our system.

We use the neural network in order to approximate the reconstructed values of the fusion center. Simulations show (see Table 3) that the neural network with enough parameters (that correspond to weights and biases of neurons) can achieve better performance than the one given by the direct sum estimation method. Moreover, with the neural network we can have control on the number of parameters that need to be stored in order to approximate the fusion center table h.

Let L_1 be the number of neurons in the first layer of the neural network. We use only one neuron for the second layer (the output layer). Let W_1, B_1, W_2, B_2 be the weight and bias matrices for the first and the second layer respectively. We use the notation $c_k : k = 1, \ldots, M$ for the elements of these weight and bias matrices that are, in other words, the parameters of the neural network. Let I be the number of inputs of the neural network. These correspond to the codewords of the quantizers. We use the unary representation of the inputs where the number of inputs is twice (for two quantizers) the total number of codewords. The dimensions for the weight and bias matrices are $L_1 \times I$, $L_1 \times 1$, $1 \times L_1$, and 1×1 respectively. Then the number of parameters, M, used for the description of the two-layer neural network is

$$L_1(I+2)+1. \tag{17}$$

The number of parameters for the weights and the biases of the neural network depends on the representation of the input. For the training we can use either the quantities, $\theta^{(t)}$, or the corresponding fusion table entry, $h^{(t)}$, given by

$$h^{(t)} = \sum_{i=1}^{N}\sum_{j=1}^{L} h(i,j) I_{U_i}(X_1^{q,t}) I_{V_j}(X_2^{r,t}) \tag{18}$$

for every point t of the training set, where $h(i,j)$ is given by Equation 6. This is because the corresponding two expressions for the estimation error differ by a quantity that depends only on the training set (the proof is easy and is omitted). However, the learning process is faster for the second case because of the variation of the target values for the same input in the first case.

If $f(i,j)$ is the output of the neural network for inputs i,j after the training, then the approximation of the fusion center table entry $h^{(t)}$, $\hat{h}^{(t)}$, is given by

$$\hat{h}^{(t)} = \sum_{i=1}^{N}\sum_{j=1}^{L} f(i,j) I_{U_i}(X_1^{q,t}) I_{V_j}(X_2^{r,t}) \tag{19}$$

for every point t of the training set. Then the total estimation error can be expressed as follows

$$Error_{NN} = \frac{1}{M}\sum_{t=1}^{M}\left(\theta^{(t)} - \hat{h}^{(t)}\right)^2 \tag{20}$$

σ_n^2	bp_init, 16 leaves, (8,8) labels					
	$\rho = 0$		$\rho = 0.5$		$\rho = 0.85$	
	no-constr. 64 params	dir.sum.est. 16 params	no-constr. 64 params	dir.sum.est. 16 params	no-constr. 64 params	dir.sum.est. 16 params
0.001	0.0058	0.0145	0.0058	0.0132	0.0063	0.0169
0.010	0.0135	0.0221	0.0145	0.0226	0.0250	0.0268
0.050	0.0381	0.0466	0.0493	0.0562	0.0554	0.0597
0.100	0.0674	0.0702	0.0878	0.0942	0.0976	0.1086
0.200	0.1154	0.1131	0.1540	0.1566	0.1816	0.1839
0.300	0.1573	0.1524	0.2013	0.2070	0.2436	0.2485
0.400	0.1897	0.1935	0.2550	0.2569	0.2998	0.3057
0.500	0.2209	0.2218	0.2945	0.2997	0.3470	0.3487
0.600	0.2521	0.2542	0.3383	0.3452	0.3901	0.3948
0.700	0.2849	0.2873	0.3730	0.3800	0.4260	0.4264
0.800	0.3126	0.3143	0.3989	0.4057	0.4643	0.4656
0.900	0.3410	0.3442	0.4234	0.4287	0.4923	0.5060
1.000	0.3621	0.3662	0.4593	0.4602	0.5126	0.5262

Table 1: Comparison of the performance of the unconstrained (full fusion center table) and the direct sum estimation approach.

In order to improve the performance of the system further we allow some changes of the labels of the quantizer rectangles taking into account the last form of the fusion center that now depends on the parameters, c's, of the neural network. We apply the lh-s-CGLA (lookahead set CGLA) in order to decide about the best label of a rectangle. The algorithm, *design-fc-q*, that iteratively trains the neural network for the fusion center and possibly alters the quantizers by assigning new labels to the rectangles is described below. The variable l is the iteration counter, h^l the fusion center at iteration l, and c^l's the neural network parameters at iteration l.

```
design-fc-q :
1. $l \leftarrow 1$.
2. run lh-s-CGLA to assign better labels to the rectangles of the quantizers.
3. calculate $h^l$.
4. calculate $c^l$'s by constructing a neural network trying to approximate
   the fusion center $h^l$.
5. calculate the estimation error given by Equation 20.
6. if the reduction of the estimation error is less than a given threshold, stop,
   else $l \leftarrow l + 1$ and go to step 2.
```

Notice that this is an attempt to partially redesign the quantizers based on the new fusion center. An approach that tries to change also the split points of the quantizers is not feasible because of the extra cost that the training of the neural network involves every time it is trying to decide where to put the split points.

5 Simulation results and discussion

In the simulations we consider the case where the observations at the quantizers are scalar quantities of the form:

$$x_k = \theta + n_k, \; k = 1, 2 \tag{21}$$

σ_n^2	16 leaves, $(8,8)$ labels bp_init, $\rho = 0.85$	
	dir.sum.est. (a)	dir.sum.est. (b)
0.001	0.2058	0.0169
0.010	0.2741	0.0268
0.050	0.1926	0.0597
0.100	0.2108	0.1086
0.200	0.2791	0.1839
0.300	0.3014	0.2485
0.400	0.3643	0.3057
0.500	0.3719	0.3487
0.600	0.4710	0.3948
0.700	0.4833	0.4264
0.800	0.5006	0.4656
0.900	0.5478	0.5060
1.000	0.5804	0.5262

Table 2: Performance improvement when we build the quantizers taking into account the special form of the fusion center function in the direct sum estimation method.

where the noises n_k at the sensors are Gaussian distributed with correlation coefficient ρ and marginal distributions $N(0, \sigma_n^2)$, where σ_n^2 is the variance of the noises. The parameter θ has Gaussian distribution $N(0,1)$ and is independent of the noises n_k, $k = 1, 2$. The quantizers are designed using a training set $\mathcal{T}$ of 5,000 samples and are tested on a test set $\mathcal{T}'$ of 5,000 samples that is independent from $\mathcal{T}$ although it is constructed the same way as $\mathcal{T}$. The results that we report here are on the test set $\mathcal{T}'$. We use the breakpoint initialization of labels. The value 0.005 was used for the error threshold in all experiments. The number of epochs that were used to train the neural network is 10,000.

In Table 1 we compare the estimation error of the unconstrained method (without restriction on the fusion center) with that of the direct sum estimation method that constrains the capabilities of the fusion center using Equation 13. We assume breakpoint initialization of labels and we present results for several values of σ_n^2 and for $\rho = 0, 0.5, 0.85$. Notice that for high noise variance the performance of the direct sum estimation system is very close to the performance of the unconstrained system.

In Table 2 we present the performance of the direct sum estimation method in the case where we do not take into account the restrictions on the fusion center when building the regression trees and we only apply the procedure that calculates the parameters of the fusion center as the last step (a). We also present the performance in the case where the procedure that builds the quantizers takes into account the special form that the fusion center function takes in order to satisfy the storage constraints (b). We present the results for $\rho = 0.85$. The difference in the estimation error is similar for $\rho = 0, 0.5$.

In Table 3 we compare the performance of the direct sum estimation system with that of the neural network with 3 and 4 neurons. The number of parameters used by each system is also shown. It is clear that the neural network approach provides flexible trade-offs between storage complexity of the fusion center (i.e. number of neurons used) and performance of the quantizers. The results that we report here for the neural network are obtained after we apply the *design-fc-q* algorithm that iteratively trains the neural network for the fusion center and possibly alters the quantizers by assigning new labels to

σ_n^2	bp_init, $\rho = 0.85$, 16 leaves, $(8,8)$ labels		
	dir.sum.est. 16 params	3 neurons 37 params	4 neurons 55 params
0.001	0.0169	0.0132	0.0064
0.010	0.0268	0.0270	0.0261
0.050	0.0597	0.0588	0.0571
0.100	0.1086	0.1078	0.0984
0.200	0.1839	0.1830	0.1819
0.300	0.2485	0.2501	0.2481
0.400	0.3057	0.2999	0.2996
0.500	0.3487	0.3480	0.3474
0.600	0.3948	0.3925	0.3920
0.700	0.4264	0.4267	0.4261
0.800	0.4656	0.4650	0.4637
0.900	0.5060	0.4982	0.4939
1.000	0.5262	0.5176	0.5144

Table 3: Using a neural network to approximate the fusion center table.

the rectangles of their regression trees. We observed that using the *design-fc-q* algorithm instead of the one-iteration approach that builds the neural network and stops, helps to improve the performance of the system.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees.* Wadsworth Inc., 1984.

[2] P. A. Chou, T. Lookabaugh, and R. M. Gray. Optimal Pruning with Applications to Tree-Structured Source Coding and Modeling. *IEEE Transactions on Information Theory*, IT-35(2):299–315, Mar. 1989.

[3] J. A. Gubner. Distributed Estimation and Quantization. *IEEE Transactions on Information Theory*, IT-39(4):1456–1459, Jul. 1993.

[4] S.-Z. Kiang, R. L. Baker, G. J. Sullivan, and C.-Y. Chiu. Recursive Optimal Pruning with Applications to Tree Structured Vector Quantizers. *IEEE Transactions on Image Processing*, 1(2):162–169, Apr. 1992.

[5] W.-M. Lam and A. R. Reibman. Quantizer design for decentralized estimation systems with communications constraints. In *Proceedings of the 23rd Annual Conference on Information Sciences and Systems, Baltimore, Maryland*, pages 489–494, Mar. 1989.

[6] W.-M. Lam and A. R. Reibman. Design of Quantizers for Decentralized Estimation Systems. *IEEE Transactions on Communications*, 41(11):1602–1605, Nov. 1993.

[7] V. Megalooikonomou and Y. Yesha. Quantization for Distributed Estimation with Unknown Observation Statistics. In *Proceedings of the 31th Annual Conference on Information Sciences and Systems, Baltimore, Maryland*, pages 138–143, Mar. 1997.

Orientation Estimation Bounds for High Resolution Radar Data

Joseph A. O'Sullivan
Vikas Kedia
Department of Electrical Engineering
Washington University
St. Louis, MO 63130

Steven P. Jacobs
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, PA 15261

September 26, 1997

Abstract

Automatic target recognition algorithms typically are based on joint inference of the target orientation and the target type. In such situations, the ability to estimate target orientation is a prerequisite for recognition, and lower bounds on orientation estimation performance imply bounds on recognition. This paper determines Hilbert-Schmidt bounds on orientation estimates for targets when the available data are from high resolution radar sensors, following the methodology and analysis of Grenander, Miller, and Srivastava. These bounds may be viewed as the extension of Cramer-Rao bounds to matrix groups. We describe a special case of these bounds for targets that exhibit physical symmetry so that orientation estimation is reduced to discriminating between equivalence classes. The bounds are computed for several targets assuming data from each of a few frequency bands. The Hilbert-Schmidt bound is attained for the Hilbert-Schmidt estimator, the equivalent of a minimum mean-square error estimator for matrix groups. For some cases, this estimator may be computationally complex, and other estimators may be desirable. We also compute the performance of the maximum a posteriori (MAP) estimator for several cases of interest.

1 Introduction

The problem of automatic target recognition often requires performing inference on target position and orientation (pose) in addition to target type. In such situations, the quality of the estimate of pose affects the probability of correct recognition. Recently, Grenander, Miller, and Srivastava [1, 4] have proposed the use of a Hilbert-Schmidt bound for predicting the performance of orientation estimation. The target is represented via a rigid body model, equipped with a group action. The group action consists of target translation and rotation. By mapping this group to its equivalent matrix group representation, a norm is induced on the group via a matrix norm. The group equivalent of a Cramer-Rao bound has been dubbed the Hilbert-Schmidt bound and is derived using this approach. We summarize the derivation, concentrating on orientation estimation, and describe a computation of these bounds for high resolution radar (HRR) targets. The concept is extended to symmetric targets so that distances are computed between equivalence classes of orientations. For HRR sensors, a likelihood model is described, based on the use of XPATCH as a target signature simulator. The computation of the bounds is applied to several XPATCH data sets, and variations on the proposed model are explored in terms of the resulting Hilbert-Schmidt bounds on target orientation estimation.

2 Hilbert-Schmidt Bounds

In typical automatic target recognition scenarios, a rigid target has an orientation at a fixed time that is in $\mathbf{SO}(3)$, the special orthogonal group of dimension 3. For ground-based targets on a flat, known ground plane, the orientation is in $\mathbf{SO}(2)$. We use an $n \times n$ rotation matrix representation for $\mathbf{SO}(n)$. If A, $B \in \mathbf{SO}(n)$, then the squared Hilbert-Schmidt distance is

$$d_{HS}(A, B) = ||A - B||_{HS}^2 = tr[(A-B)^T(A-B)], \tag{1}$$

where $tr[\cdot]$ is the matrix trace operation and $(\cdot)^T$ denotes transpose. Since $tr[A^TA] = n$ for all $A \in \mathbf{SO}(n)$,

$$d_{HS}(A, B) = 2n - 2tr[A^TB] = 2n - 2tr[AB^T]. \tag{2}$$

For ground-based targets, a 2×2 rotation matrix $\mathbf{O}(\theta) \in \mathbf{SO}(2)$ defines the orientation of a target relative to a high resolution radar. For example, let $\mathbf{O}$ define the orientation of a ground target illuminated by an airborne radar platform. For a fixed elevation angle, θ is the azimuth of the target relative to the radar. Because $\mathbf{O}$ is a rotation matrix,

$$\mathbf{O}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{3}$$

and for any two rotations $\mathbf{O}_1$ and $\mathbf{O}_2$ [1, 4],

$$d_{HS}(\mathbf{O}_1, \mathbf{O}_2) = 4 - 4\cos(\theta_1 - \theta_2). \tag{4}$$

Given a true orientation for the target $\mathbf{O}_{true}$ and any estimate $\hat{\mathbf{O}}$, the squared Hilbert-Schmidt norm $||\mathbf{O}_{true} - \hat{\mathbf{O}}||_{HS}^2$ gives a measure of the squared error in the estimate. Let $\mathbf{r}$ denote an observed range-profile from which the estimate $\hat{\mathbf{O}}$ is computed, and let $p(\mathbf{r}|\mathbf{O})$ denote the likelihood function for $\mathbf{r}$ given $\mathbf{O}$. Also, let $p(\mathbf{O})$ be the prior on the orientation. A Hilbert-Schmidt estimator is defined as an estimator that, given $\mathbf{r}$, minimizes the conditional mean under the posterior of the squared estimation error as defined by the Hilbert-Schmidt norm squared [4],

$$\hat{\mathbf{O}}_{HS}(\mathbf{r}) = \underset{\mathbf{O} \in \mathbf{SO}(2)}{\operatorname{argmin}} \int_{\mathbf{SO}(2)} ||\mathbf{O} - \mathbf{O}'||_{HS}^2 p(\mathbf{O}'|\mathbf{r}) \gamma(d\mathbf{O}'), \tag{5}$$

where the posterior $p(\mathbf{O}|\mathbf{r}) \propto p(\mathbf{r}|\mathbf{O})p(\mathbf{O})$ and γ is the base measure on $\mathbf{SO}(2)$. Note that this is a minimum mean-squared error estimator, defined in terms of the Hilbert-Schmidt norm [1, 4].

The Hilbert-Schmidt bound on the expected squared error in estimation of a rotation in $\mathbf{SO}(2)$ is defined as [1, 4]

$$HSB = E_{\mathbf{O}}\left\{E_{\mathbf{r}|\mathbf{O}}\left\{||\hat{\mathbf{O}}_{HS}(\mathbf{r}) - \mathbf{O}||_{HS}^2 \mid \mathbf{O}\right\}\right\}, \tag{6}$$

where the inner expectation is over the ensemble of observations $\mathbf{r}$ given the orientation, and the outer expectation is over the set of possible true orientations $\mathbf{O} \in \mathbf{SO}(2)$.

2.1 Target-Sensor Symmetry

For a specific combination of target and sensor, symmetries in the target shape or the physics of the sensor may lead to equivalences in the sensor output for an entire class of target orientations. This in turn results in fundamental ambiguities when attempting to infer the orientation of the target from sensor data. Suppose that a likelihood model for data given a target and the target orientation is given by $p(\mathbf{r}|\mathbf{O}, a)$. Target symmetry is defined for a specific sensor, and that symmetry induces subsets of the orientations over which the likelihood function is constant.

Definition: A target is *symmetric* with respect to a sensor and a set of equivalence classes $\{\Phi(\psi), \psi \in \Psi\}$ if

$$\cup_{\psi \in \Psi} \Phi(\psi) = \mathbf{SO}(n), \tag{7}$$

and

$$p(\mathbf{r}|\mathbf{O}_1, a) = p(\mathbf{r}|\mathbf{O}_2, a), \tag{8}$$

for all $\mathbf{O}_1, \mathbf{O}_2 \in \Phi(\psi)$.

That is, the equivalence classes form a partition of $\mathbf{SO}(n)$ and the likelihood function is constant over an equivalence class. The induced Hilbert-Schmidt distance that is appropriate in this case is between equivalence classes:

$$d_{HS}(\psi_1, \psi_2) = \min_{\mathbf{O}_1 \in \Phi(\psi_1), \mathbf{O}_2 \in \Phi(\psi_2)} d_{HS}(\mathbf{O}_1, \mathbf{O}_2). \tag{9}$$

Many targets are left-right symmetric with respect to HRR range profile sensors. If the radar transmitter emits waves with circular polarization, then all targets are symmetric with respect to the corresponding receiver with equivalence classes consisting of rotations around the line-of-sight (see [2]).

2.2 Computing Hilbert-Schmidt Bounds

Computation of the Hilbert-Schmidt bound involves evaluating the two expectations given in (6). The inner expectation is approximated by averaging over multiple realizations of the noise added to form the observations. The outer expectation is approximated by averaging over a uniform sampling of orientations in $\mathbf{SO}(2)$. For the purposes of computing the bound, we assume a uniform prior $p(\mathbf{O})$. Under these conditions, the posterior is determined by the data likelihood.

For a fixed true orientation $\mathbf{O}$ and a given realization of the data $\mathbf{r}$, the Hilbert-Schmidt estimate is computed as follows. The matrix $\mathbf{A}$ is defined as the conditional mean of $\mathbf{O}$ under the posterior $p(\mathbf{O}|\mathbf{r})$,

$$\mathbf{A} = \int_{\mathbf{SO}(2)} \mathbf{O}' p(\mathbf{O}'|\mathbf{r}) \gamma(d\mathbf{O}'). \tag{10}$$

This matrix is approximated by computing the sample average $\tilde{\mathbf{A}}$,

$$\tilde{\mathbf{A}} = \frac{\sum_{i=0}^{N-1} \mathbf{O}'(\theta_i) p(\mathbf{r}|\mathbf{O}'(\theta_i))}{\sum_{i=0}^{N-1} p(\mathbf{r}|\mathbf{O}'(\theta_i))}, \tag{11}$$

where the angles θ_i are equally spaced around the circle, $\theta_i = \frac{2\pi i}{N}$. The Hilbert-Schmidt estimate is found by normalizing $\tilde{\mathbf{A}}$ to unit determinant, which projects this matrix into $\mathbf{SO}(2)$,

$$\hat{\mathbf{O}}_{HS} = \frac{\tilde{\mathbf{A}}}{\sqrt{|\tilde{\mathbf{A}}|}}. \tag{12}$$

For a fixed true orientation $\mathbf{O}(\theta_i)$, the inner expectation in (6) is computed by generating M independent realizations of the data $\mathbf{r}$ according to $p(\mathbf{r}|\mathbf{O}(\theta_i))$. For each realization $\mathbf{r}_k$, the Hilbert-Schmidt estimate $\hat{\mathbf{O}}_{HS}(\mathbf{r}_k)$ is computed as in (12), and the Hilbert-Schmidt distance $d_{HS}(\mathbf{O}(\theta_i), \hat{\mathbf{O}}_{HS}(\mathbf{r}_k))$ is accumulated. The inner expectation is approximated by

$$\sum_{k=1}^{M} \frac{1}{M} d_{HS}(\mathbf{O}(\theta_i), \hat{\mathbf{O}}_{HS}(\mathbf{r}_k)). \tag{13}$$

The above calculations are repeated for N true orientations $\mathbf{O}(\theta_i)$ where $\theta_i = \frac{2\pi i}{N}$, and the HSB is found by averaging the results over the N true orientations. Finally, the entire process is repeated for different noise variances N_0, so that we may examine the variation in the HSB with SNR for a particular sensor.

3 HRR Range Profile Likelihoods

A crucial issue relating to orientation estimation and target recognition using a radar sensor is the modeling of HRR range-profiles as represented by the likelihood function for the data, in this case $p(\mathbf{r}|\mathbf{O})$. In this paper, we make exclusive use of a deterministic model for the range-profiles. The received data are given by

$$r(t) = s(t; \mathbf{O}, a) + w(t), \tag{14}$$

where $s(t; \mathbf{O}, a)$ is a deterministic signal term that is dependent on both the target type $a \in \mathcal{A}$ and orientation $\mathbf{O}$, and $w(t)$ is additive complex white Gaussian noise. The data actually available are a discretized version of the random process $r(t)$ formed by integration against orthonormal functions,

$$\mathbf{r} = \mathbf{s}(\mathbf{O}, a) + \mathbf{w}. \tag{15}$$

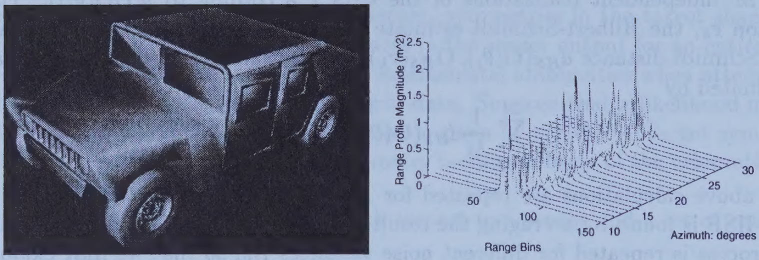
The entries of $\mathbf{w}$ are then independent complex Gaussian random variables with mean 0 and variance N_0, and $\mathbf{s}(\mathbf{O}, a)$ is a deterministic signal vector as produced, for example, by an appropriate signature prediction tool. In this case, the log-likelihood for $\mathbf{r}$ is

$$L(\mathbf{r}|\mathbf{O}, a) = \frac{2}{N_0} Re\left\{ \left[\mathbf{r} - \frac{1}{2}\mathbf{s}(\mathbf{O}, a)\right]^{\dagger} \mathbf{s}(\mathbf{O}, a) \right\}, \tag{16}$$

where $[\cdot]^{\dagger}$ denotes the Hermitian transpose. The dependence of $\mathbf{s}$ on a is crucial from the point of view of target recognition, but in evaluating bounds on orientation estimation we will hold the target type constant, and the dependency is ignored. In related work [3], we consider the relative merits of deterministic and stochastic models for the HRR data, and introduce a conditionally Gaussian model that may be more appropriate for ATR than the deterministic model used here.

4 Range Profile Simulation

All of the high resolution radar range-profiles used in this paper are simulated data computed using a HRR signature prediction tool known as XPATCH. Developed jointly by Wright Laboratory and DEMACO, Inc., XPATCH uses a ray tracing algorithm and a CAD model for the target to simulate the range-profile. The user specifies the frequency



HMMWV CAD model Simulated Range Profiles

Figure 1: CAD model of the HMMWV and millimeter-wave range-profiles generated using XPATCH.

band in which the sensor operates, and the orientation of the target relative to the sensor. Figure 1 displays a CAD model for a HMMWV ground vehicle and several millimeter-wave range-profiles produced by using XPATCH with this vehicle at different azimuth angles. By collecting such data for a fixed elevation and a uniform sampling of the possible azimuths, we obtain a data set appropriate for approximating the Hilbert-Schmidt bound given by (6).

4.1 Left-Right Symmetric Targets

Examination of the CAD model in Figure 1 reveals that the HMMWV, like many aircraft and ground vehicles, exhibits left-right or y-symmetry. Assume that the reflectivity of the target is given by the summed contribution of a finite number of scattering sites, and the range-profile results from projection of the individual reflectivities along the line of sight and sorting into range bins. Under these conditions the range-profile will be invariant to reflection in the plane of symmetry. In other words, a symmetric target at an azimuth angle of θ will have the same distribution of reflectivity along the line-of-sight as the same target at an azimuth angle of $-\theta$. This establishes a set of equivalence classes, each containing two distinct orientations that yield identical range-profiles.

The effect of left-right symmetry on orientation estimation is demonstrated by Figure 2. In this figure, the log-likelihood in (16) is plotted as a function of the true azimuth angle (x-axis) and the hypothesized azimuth angle (y-axis). The log-likelihood has been normalized by the maximum in each column, and negative values have been set to zero. The characteristic "X" pattern seen along the diagonals is indicative of the symmetry present in the range-profiles. The implications of this symmetry are that orientation estimation is possible only to the level of the equivalence class without additional information such as tracking data, and that in computing the Hilbert-Schmidt distance between two orientations in (6) we must find the minimum distance between all four elements of the associated equivalence classes, as in (9). We have effectively implemented this strategy when computing Hilbert-Schmidt bounds for symmetric targets by restricting the possible azimuth angles to lie between 0 and 180 degrees.

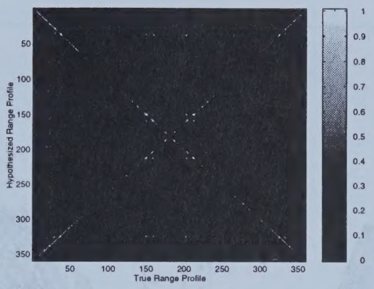


Figure 2: Log-likelihood of observing a particular range profile (y-axis) given the underlying true orientation (x-axis), for the HMMWV using mm-wave data. The log-likelihood has been normalized by the maximum in each column. Negative values have been made 0. The color at each pixel indicates its value.

4.2 Hilbert-Schmidt Bound Computations

Figure 3 shows the results of computing the inner expectation in (6) as described above, and repeating the computation for true azimuth angles between 1° and 180°, and for signal-to-noise ratios between -10 dB and 30 dB. The figure shows the results using simulated mm-wave, L-band and X-band range-profiles for the HMMWV target. In the left column of the figure, orientation estimates were obtained using the minimum mean-squared error (MMSE) estimator given in (5). In the right column, a maximum a-posteriori probability (MAP) estimator was used. Because the prior $p(\mathbf{O})$ is uniform, this estimator simply determines the azimuth angle that maximizes the deterministic model log-likelihood given in (16). Note that in all cases, these computations indicate larger errors in orientation estimates when the underlying true orientation is near head-on or tail-on than when it is near broadside.

Figure 4 shows the Hilbert-Schmidt Bound obtained by averaging the results in Figure 3 over all true azimuth angles. The left panel shows the HSB using the MMSE estimator, and the right panel shows the result using the MAP estimator. In this figure, a value of 0 for the HSB indicates perfect azimuth estimation, in that the estimate is consistently equal to the truth. A value of $4 - \frac{8}{\pi} \approx 1.45$ indicates completely unreliable azimuth estimation, in the sense that this is the performance that would be achieved by an estimator whose estimates are uniformly distributed between 0 and 180 degrees. Since the MMSE (Hilbert-Schmidt) estimator is optimal for this performance criterion, the performance of the MMSE estimator is superior to that of the MAP estimator for all frequency bands and at all noise variances. For SNR greater than 20 dB, the MAP estimator has performance comparable to the MMSE estimator. For both the MMSE and MAP estimators, the performance is no better than the worst-case estimator for SNR values less than 10 dB, and that the performance improves quickly as the SNR is increased above 10 dB. In fact, the performance of the MAP estimator is significantly worse than the uniform estimator at low SNR. This occurs because the signal energy for range-profiles at head-on, broadside, and tail-on orientations is much larger than that at other orientations. As a result, at low SNR the MAP estimator chooses one of these three orientations as the estimate with high probability, regardless of the true orientation.

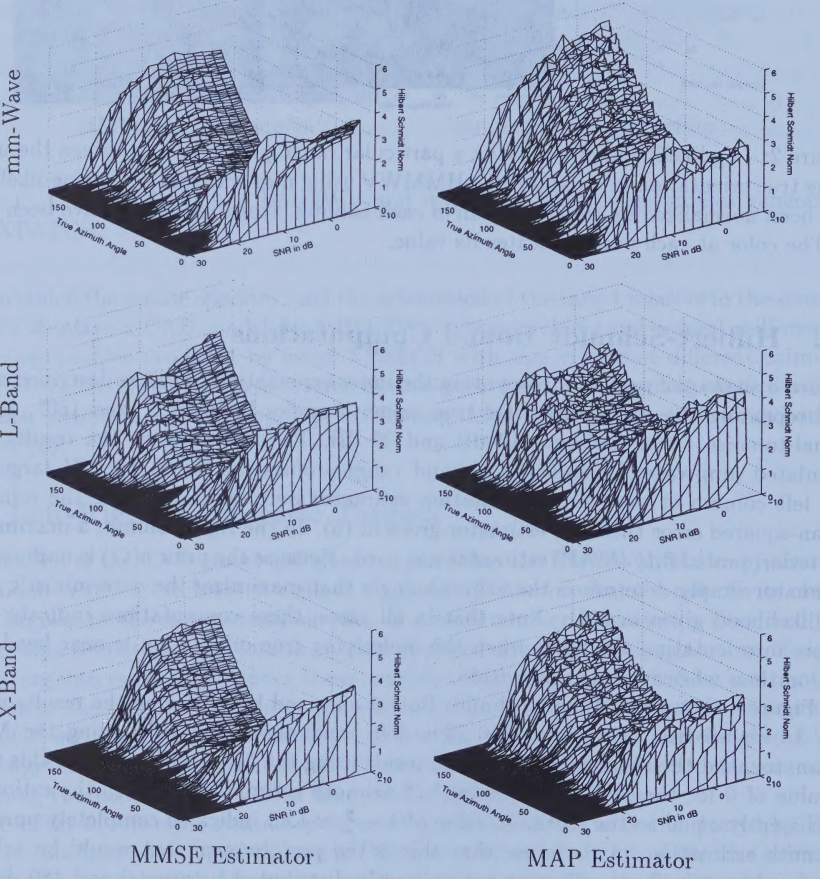
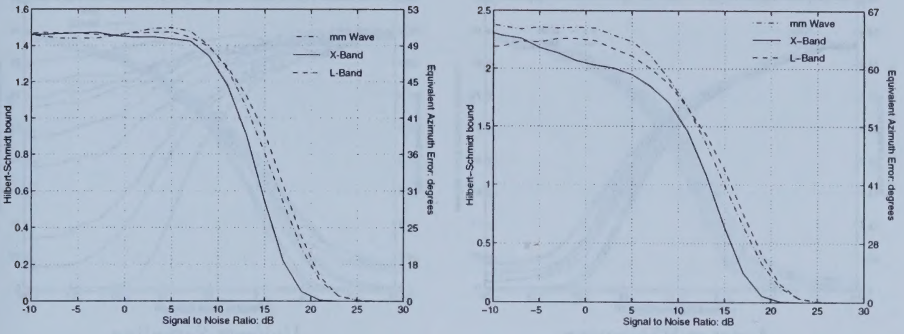


Figure 3: Average Hilbert-Schmidt norm-squared computed using mm-Wave, L-Band and X-Band data for the HMMWV target, as a function of the true azimuth angle and the SNR of the observations. Results shown are for the minimum mean-squared error (MMSE) estimator in the left column, and the maximum a-posteriori probability (MAP) estimator in the right column.



Average, MMSE Average, MAP

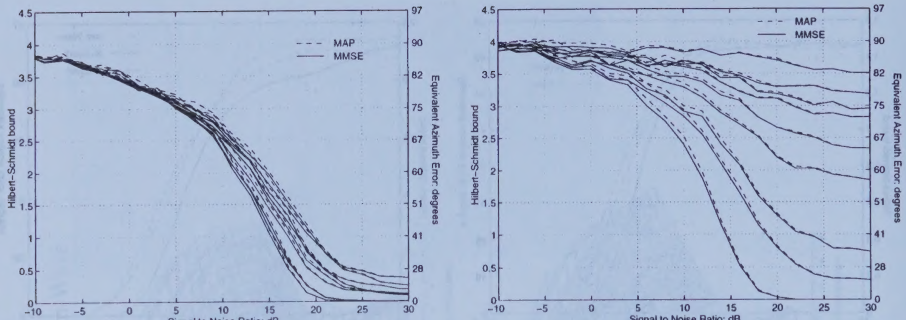
Figure 4: Hilbert-Schmidt Bounds averaged over all azimuth angles for mm-Wave, L-Band and X-Band data using MMSE and MAP estimators.

4.3 HSB for Asymmetric Targets

In addition to the simulated HRR data described earlier, we have made use of a standardized collection of XPATCH range-profiles known as the University Research Initiative Synthetic Dataset or URISD. The URISD contains range-profiles of four different vehicles, for three frequency bands, at three different elevation angles, and for a uniform sampling of the possible azimuth angles. The CAD models used in producing the URISD contain a simulated rough ground surface, included so that the data exhibit the effects of ground clutter in the vicinity of the target. Unlike the vehicles themselves, the ground surface used is not physically symmetric. Therefore, the range-profiles in the URISD do not have the symmetry property indicated in Figure 2, and when computing bounds with these data we have included candidate azimuth angles over the entire circle.

The result of computing the Hilbert-Schmidt bound using UHF-band range-profiles for a school bus from the URISD is displayed in Figure 5. Because we are now performing azimuth estimation over the entire circle, rather than over only 180 degrees as in Figure 4, the worst-case estimator yields a Hilbert-Schmidt bound value of 4. Both MMSE and MAP estimators produce nearly worst-case errors when the SNR is low, and the performance improves rapidly for SNR above 10 dB. The MAP and the MMSE estimators have comparable performance for high SNR values, and the performance degradation for using the MAP estimator is small at lower SNR.

The URISD contains a very large number of range-profiles for each target. We have therefore implemented two techniques to reduce the data storage and computation time necessary to compute the bounds: average windowing and uniform sampling. In average windowing, a new dataset is made by averaging n consecutive range-profiles from the original URISD dataset. In uniform sampling, only every nth range-profile from the URISD dataset is used. In both cases, data storage and computation time reduce by a factor of n. Figure 5 also shows the results of these two techniques for values of n between 1 and 9. Note that in both cases, the best performance is achieved for $n = 1$ (i.e., when every range-profile is used) and the performance degrades as n is increased. However, it is seen clearly that the degradation in performance is much less severe with



Average Windowing Uniform Sampling

Figure 5: Hilbert-Schmidt bound using UHF URISD data for a school bus. In average windowing, the dictionary consists of range-profiles generated by averaging n consecutive range-profiles from the URISD data set. In uniform sampling, every nth range-profile is used as an element in the dictionary. Results shown are for $n = 1, \cdots, 9$.

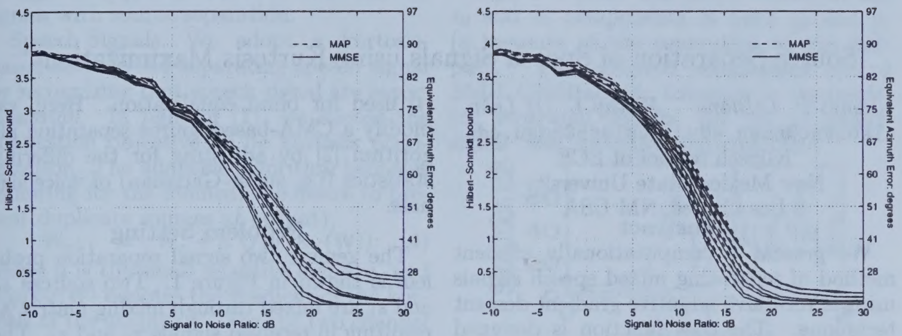
average windowing than with uniform sampling. Figure 6 shows the results using URISD UHF-band range-profiles for a fire truck and two tank models known simply as M1 and T1. The results are shown for average windowing for the window size n increasing from 1 to 9, and are comparable to those seen in the left panel of Figure 5. We are presently extending this approach to an adaptive windowing technique, in which the size of the window varies depending on the correlation among neighboring range-profiles.

5 Conclusions

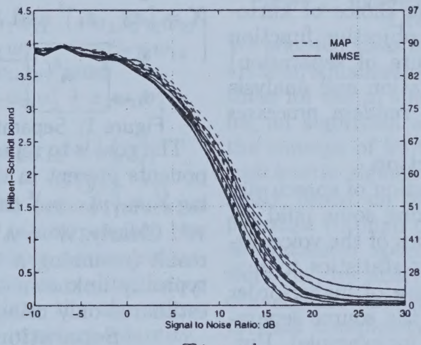
Accurate recognition of targets often requires high quality estimates of target orientation. Hilbert-Schmidt bounds for estimating target orientation have been presented. These bounds have been extended to account for target symmetry, and bounds for discriminating between equivalence classes induced by target symmetry have been derived. Computation of the bounds has been performed using a variety of simulated datasets, yielding results consistent with the optimality of the Hilbert-Schmidt estimator. For high SNR values, the MAP estimator has performance comparable to the Hilbert-Schmidt (MMSE) estimator. Extensive use has been made of the radar signature prediction tool XPATCH, including use of the URISD.

Appropriate models for targets have been explored in a limited way, focusing on variations of a deterministic model. In any given scenario, the computational resources are generally limited, so use of a database the size of the URISD is generally not feasible. Subsampling this database is shown here to give inferior results as compared to an averaging approach. Several other variations on this approach are being actively pursued, including the use of various conditionally Gaussian models.

Acknowledgements. This work supported in part by the ARO grant for the Center for Imaging Science DAAH049510494.



Fire Truck M1 tank



T1 tank

Figure 6: Hilbert-Schmidt bound using UHF URISD data for fire truck, M1 tank, and T1 tank. Results shown are for average windowing with $n = 1, \cdots, 9$

References

[1] U. Grenander, M. I. Miller, and A. Srivastava. Hilbert-schmidt lower bounds for estimators on matrix lie groups. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 1997. Submitted for publication.

[2] S. P. Jacobs. *Automatic Target Recognition Using High Resolution Radar Range-Profiles.* PhD thesis, Washington University, Saint Louis, Missouri, May 1997.

[3] S. P. Jacobs and J. A. O'Sullivan. Automatic target recognition using sequences of high resolution radar range-profiles. *IEEE Transactions on Aerospace and Electronic Systems.* Submitted for publication June 17, 1997.

[4] A. Srivastava. *Inferences on Transformation Groups Generating Patterns on Rigid Motions.* PhD thesis, Washington University, Saint Louis, Missouri, July 1996.

Source Separation of Speech Signals using Kurtosis Maximization

James P. LeBlanc *Phillip L. De Leòn*
`leblanc@nmsu.edu` `pdeleon@nmsu.edu`
Klipsch School of ECE
New Mexico State University
Las Cruces, NM USA

Abstract

We present a computationally efficient method of separating mixed speech signals using a recursive adaptive gradient descent technique. The cost function is designed to maximize the kurtosis of the output (separated) signals. The choice of kurtosis maximization as an objective function (which acts as a measure of separation) is supported by investigation and analysis of *spherically invariant random processes* (SIRP's) [6].

Introduction

The problem of separation of speech signals is considered. Making some mild assumptions on the statistics of the voice signals we use higher order statistics to separate the voices. The use of higher-order statistics is not new to the source separation problem (see [1], [5], for example). But, many of these methods are applied to digital communications signals which belong to a different statistical class (e. g. sub-Gaussian) [1] than speech signals (super-Gaussian).

A fundamental idea of many blind separation and equalization schemes in digital communications makes note that the sum of sub-Gaussian processes (as occurs with mixing and intersymbol interference) results in a process that "looks more" Gaussian than the originals. [3] includes an excellent discussion of measures of *Gaussianity.* With such a measure, one constructs a cost function, and associated adaptive gradient descent algorithm which minimizes this Gaussianity measure resulting in source separation or intersymbol interference reduction. A common measure which appears quite often is *kurtosis*, which is defined for a zero mean random process X as $\kappa_X = \mathrm{E}\{x^4\}/\{\mathrm{E}\{x^2\}\}^2$ Kurtosis relates to the Constant Modulus Algorithm (CMA) [4] used for blind equalization. Here, we modify a CMA-based source separation algorithm [2] by adjusting for the differing statistics (i.e. super-Gaussian) of voice signals.

Problem Setting

The generic two signal separation problem is shown in Figure 1. Two sources s_0 and s_1 are mixed through mixing matrix $\mathbb{A}$, resulting in received signals x_0 and x_1. The mixing relation is denoted, $X = \mathbb{A}S$ where $X = [\,x_0 \quad x_1\,]^t$ and $S = [\,s_0 \quad s_1\,]^t$.

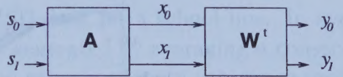


Figure 1: Separation Block Diagram

The goal is to separate the s_0 and s_1 components present in the mixed signals yielding $Y = [\,y_0 \quad y_1\,]^t$ through the use of matrix $\mathbb{W}^t$. Clearly, $\mathbb{W}^t = \mathbb{A}^{t^{-1}}$ achieves the desired result (assuming $\mathbb{A}$ is invertible) but, $\mathbb{A}$ is typically unknown and $\mathbb{W}$ (or $\mathbb{A}$) must be estimated only using the mixture X.

Separation by Kurtosis

Communications Signals. An interesting feature of kurtosis is now noted. Let u_0 and u_1 be two independent, identically distribute (iid), zero mean random variables with kurtosis κ_U. Let $w = u_0 + u_1$ and consider κ_W. It can be shown that

$$\begin{aligned} \kappa_U < 3 &\Rightarrow \kappa_W > \kappa_U \text{ (sub-Gaussian)} \\ \kappa_U > 3 &\Rightarrow \kappa_W < \kappa_U \text{ (super-Gaussian)} \end{aligned} \quad (1)$$

Since digital communications signal are typically considered to be *sub-Gaussian*, the resulting mixture will have a higher kurtosis. Thus, a logical separation strategy is to minimize the output kurtosis, which in effect, is exactly what CMA does. In [2] an iterative separation algorithm from digital communications signals utilizing the CMA error function is presented as

$$\mathbb{W}_{n+1} = \mathbb{W}_n - \mu \nabla_{\mathbb{W}} (\phi(\mathbb{W})) \quad (2)$$

where μ is the small adaptive stepsize, and $\nabla_{\mathbb{W}}\phi(\mathbb{W})$ denotes the gradient of ϕ,

$$\phi(\mathbb{W}) = \sum_{i=1}^{N} \mathrm{E}\left\{(y_i^2 - 1)^2\right\} - \ln(\det|\mathbb{W}|) \quad (3)$$

The first term is the CMA cost function, while the second term associates a cost to duplicating a source at the output Y.

---

[1] The term *sub-Gaussian* (*super-Gaussian*) is used to denote processes having a kurtosis less (more) than the kurtosis of a Gaussian.

In light of (1), such kurtosis minimization agrees with source separation.

Speech Signals. We adopt a kurtosis-based strategy for separating speech signal by recognizing that speech signal are *super-Gaussian*. In light of (1), we choose the adaptation objective to be kurtosis maximization. The adaptive algorithm becomes (ignoring for the moment the desire to prevent duplicate sources at output),

$$\mathbb{W}_{n+1} = \mathbb{W}_n + \mu \nabla_{\mathbb{W}} (\kappa_Y(\mathbb{W})) \quad (4)$$

where μ is the small adaptive stepsize, and $\nabla_{\mathbb{W}}\kappa_Y(\mathbb{W})$ denotes the gradient of the kurtosis of the outputs Y. For the two channel case, performing the differentiation leads to the update law

$$\mathbb{W}_{n+1} = \mathbb{W}_n + \mu \begin{bmatrix} -\alpha_1\beta_1\gamma_1 w_{21} & -\alpha_2\beta_2\gamma_2 w_{22} \\ \alpha_1\beta_1\gamma_1 w_{11} & \alpha_2\beta_2\gamma_2 w_{12} \end{bmatrix}$$

where $\alpha_i = 4(w_{i1}x_1 + w_{2i}x_2)^3$ and

$$\beta_i = (-x_1 w_{i1} r_{12} - x_1 w_{2i}\sigma_2^2 + x_2 w_{1i}\sigma_1^2 + w_{2i}x_2 r_{12})$$

$$\gamma_i = 1/(w_{i1}^2\sigma_1^2 + 2w_{i1}w_{2i}r_{12} + w_{2i}^2\sigma_2^2)^3$$

and $\sigma_1^2 = \mathrm{E}\{y_1^2\}$, $\sigma_2^2 = \mathrm{E}\{y_2^2\}$, $r_{12} = \mathrm{E}\{y_1y_2\}$. Knowing the actual values of σ_1^2,σ_2^2, and r_{12} a priori is not possible but may be replaced by simple autoregressive estimators. Also, an output scaling factor must be incorporated into the algorithm, since kurtosis is a scale invariant quantity.

The critical assumption here is that the kurtosis of two mixed voice signals has a lower kurtosis than the individual kurtosis values (as *hinted* at by (1)). However, this may not be true. For speech signals there is no assurance different speakers have identical distributions, nor are the samples from any speaker temporally independent. While lacking a proof, this critical assumption holds for all sampled speech we've tested. This issue has been initially investigate using SIRP's are a good statistical speech model. Under our analysis, the critical assumption holds for a wide range of parameters representing speech. While lending credence to this approach, further study is on-going.

Separation Example

The algorithm is demonstrated for

$$\mathbb{A} = \begin{bmatrix} 0.72 & 0.34 \\ 0.41 & 0.63 \end{bmatrix},$$

s_0 is a speaker with $\kappa_{s_0} = 13.9$ and s_1 is a different speaker $\kappa_{s_1} = 13.1$. The received mixtures have kurtosis $\kappa_{x_0} = 11.4$ and $\kappa_{x_1} = 8.8$ verifying the critical assumption. The severity of the mixing renders the signals x_0 and x_1 unintelligible.

Figure 2 plots the power ratios of the s_0 and s_1 components in both y_0 and y_1 (a measure of the separation at the output). The achieved separations exceed 35dB. Qualitatively, listening to the resulting separated signals, the second speech signal was virtually imperceptible.

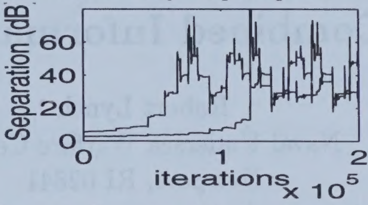


Figure 2: Example of Source Separation

Conclusion

We adopted ideas from the blind source separation/equalization problem and modified them for use in speech separation, presenting an algorithm and its motivation using the concept of kurtosis maximization with leptokurtic signals. The algorithm, originally based on heuristics, appears more plausible through use of analysis of SIRP's as statistical speech models. Continuing work on the analysis of the SIRP model, convergence analysis, and lowered computational complexity of the presented algorithm are all avenues for continuing work.

References

[1] E. Moreau, O. Macchi, "New self-adaptive algorithm for source separation based on contrast functions," *Proc. IEEE Sig. Proc. Wrkshp on Higher Order Stat.*, Lake Tahoe, CA, 1993, pp.215-219.

[2] L. Castedo, O. Macchi, "Maximizing the Information Transfer for Adaptive Unsupervised Source Separation," *Proc. IEEE Sig. Proc. Adv. in Wireless Comm. Wrkshp*, Paris, 1997, pp.65-69.

[3] D.L. Donoho, "On Minimum Entropy Deconvolution," *Applied Time Series Analysis*, D.F. Findley, Ed., New York: Academic Press, 1981.

[4] J.R. Treichler, M. G. Agee, "A New Approach to Multipath Correction of Constant Modulus Signals," *IEEE Trans. Acous., Speech, and Sig. Proc.*, Apr. 1983.

[5] J. -F. Cardoso, "Source separation using higher order moments," *Proc. ICASSP 89*, Glasgow, Scotland, May 1989, vol. 4, pp.2109-2112.

[6] Brehm, H. and Stammler, W., "Description and generation of spherically invariant speech-model signals," *Sig. Proc., vol. 12, no. 2, pp. 119-141*, Mar. 1987.

Discrete Symbol Quantity and the Minimum Probability of Error for a Combined Information Classification Test *

Robert Lynch
Naval Undersea Warfare Center
Newport, RI 02841

Peter Willett
University of Connecticut
Storrs, CT 06269

I. Introduction

In this paper, performance of the *combined Bayes* test, which *combines the information in discrete training and test data,* is studied. By "discrete" we mean that the data used to represent each class can take on one of M possible values (e.g., quantizing feature vectors into one of M symbols). In the combined Bayes test both training and test data are used to infer the true symbol probabilities while simultaneously being used in the classification decision. This test was shown in simulation, [1], to be more robust than a typical, nonadaptible, generalized likelihood ratio type test (this is referred to as the standard GLRT approach), which uses only training data to infer the true symbol probabilities.

The combined Bayes test is reproduced from [1] and it appears as

$$\frac{P\left(\vec{x}_k, \vec{x}_l, \vec{y} | \vec{y} \in k\right)}{P\left(\vec{x}_k, \vec{x}_l, \vec{y} | \vec{y} \in l\right)} = \frac{(N_l + N_{\vec{y}} + M - 1)!\,(N_k + M - 1)!}{(N_k + N_{\vec{y}} + M - 1)!\,(N_l + M - 1)!} \prod_{i=1}^{M} \frac{(x_{k,i} + y_i)!\,(x_{l,i})!}{(x_{l,i} + y_i)!\,(x_{k,i})!} \underset{l}{\overset{k}{\gtrless}} \tau \quad (1)$$

where (note, k and l are exchangeable)
$k, l \in \{\text{target, nontarget}\}$, and $k \neq l$;
M is the number of discrete symbols;
$x_{k,i}$ is the number of occurrences of the i^{th} symbol in the training data for class k;
y_i is the number of occurrences of the i^{th} symbol in the test data;
$N_k \left\{N_k = \sum_{i=1}^{M} x_{k,i}\right\}$ is the total number of occurrences of the M symbols in the training data for class k;
$N_{\vec{y}} \left\{N_{\vec{y}} = \sum_{i=1}^{M} y_i\right\}$ is the total number of occurrences of the M symbols in the test data;
τ is a decision threshold.

Notice, a key assumption behind (1) is that the same underlying symbol distribution (i.e., $\vec{p}_{k,\vec{y}\in k}$ when class k is true) produces, independently, both training and test data. This is evident in the term $(x_{k,i} + y_i)!$ where training and test data are combined.

The combined Bayes test was developed using a combined multinomial model, [1], and the Dirichlet distribution as a prior whose symbol probabilities are *uniformly distributed on the unit hyperplane.* Selection of a uniform prior is motivated by a desire to model *complete* uncertainty in knowing the true symbol probabilities of each class.

---

*This work was partially supported by a NUWC In-House Laboratory Independent Research grant.

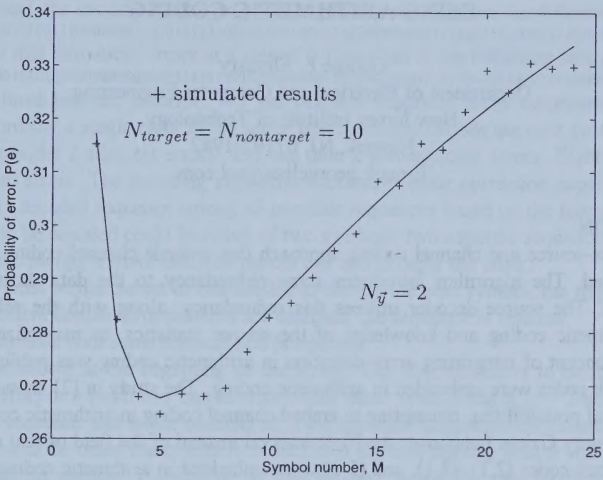


Figure 1. $P(e)$ for various numbers of discrete symbols M.

II. Results

Performance of the combined Bayes test is examined theoretically by developing a formula for its probability of error, $P(e)$, in correct classification. Figure 1 above is a plot of $P(e)$ with varying M (+ symbols indicate simulated results based on 10,000 random outcomes to verify theoretical results), where the threshold $\tau = 1$, and the priors on classes k and l are both 0.5. In this particular case $N_{target} = N_{nontarget} = 10$, and $N_{\vec{y}} = 2$. Notice that $P(e)$ starts out decreasing with increasing M reaching a minimum at which point it begins to increase with increasing M. The minimum point is referred to as M^* and it occurs at $M^* = 5$. What this curve implies is that in the absence of any prior knowledge about the discrete symbol probabilities, on average, and with ten samples of training data for each class and two test observations, best classification performance requires five discrete symbols. If more than five symbols are used, then the probability increases that a test symbol is unrepresented in the training data, and this causes more classification errors. On the other hand, if not enough discrete symbols are used then the true probabilistic structure of each class may not be adequately represented, and this also increases $P(e)$. With these results, other findings are that M^* increases, approximately linearly, when either the training or test data set size is increased. Also, in a similar analysis of the standard GLRT, it was found that M^* increases with training data set size but at a slower rate (except when $N_{\vec{y}} = 1$ in which case both tests perform the same). Additionally, when the numbers of test observations are increased M^* either stays the same, or decreases. This is attributed to the standard GLRT's inability to use test observations when inferring the symbol probabilities.

References

[1] R. Lynch and P. Willett, "Classification With a Combined Information Test," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* May 1996, pp. 3061-3064.

JOINT LOSSLESS-SOURCE AND CHANNEL CODING USING ARITHMETIC CODING

George F. Elmasry
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102-1982
E-mail: georgelisa@aol.com

INTRODUCTION

A joint lossless-source and channel coding approach that embeds channel coding within arithmetic coding is exploited. The algorithm introduces some redundancy to the data during the arithmetic encoding process. The source decoder utilizes this redundancy, along with the self-synchronization property of arithmetic coding and knowledge of the source statistics, to maximize error correction capability. The concept of integrating error detection in arithmetic coding was published in [1]. In [2] some simple block codes were embedded in arithmetic coding. The study in [2] showed that for source symbols with equal probabilities, attempting to embed channel coding in arithmetic coding is equivalent to constructing binary Galois fields over the [0-1) interval instead of the field of real numbers. Refer to [2] to see how block codes (2,1), (3,1), and (5,2) were embedded in arithmetic coding. In this paper an algorithm that embeds channel coding in arithmetic coding by keeping a certain minimum Hamming distance d_{min} between the code points assigned for each source symbol is introduced. The value of d_{min} provides a desired level of error protection. The algorithm's goal is to maximize error protection capability for compressed data. The algorithm is implemented with $d_{min} \geq 3$. Generalizing to higher or lower error protection capability is not difficult after the rules have been established. For simplicity, previous knowledge of the source statistics is assumed. Also, some rounding of the source probabilities is necessary, so no overrun or carryover is considered.

THE ALGORITHM

To find a set of code points over the [0-1) interval with $d_{min} \geq 3$ for a source alphabet with m symbols where $m = 2^k$, apply the following steps:

1- Generate list 1 of m entries, each entry being k bits long. d_{min} in this list is 1.
2- Generate list 2 by adding a parity check bit to each entry in list 1 to form m entries with $d_{min} = 2$.
3- Divide the [0-1) interval into 2^{k+1} segments and assign a segment to each symbol according to list 2.
4- Divide each of the 2^{k+1} segments into m subintervals based on the symbol's probabilities.
5- Based on the segment number and the symbol subinterval, assign code points to symbols.

Example: Suppose we have a source alphabet of 8 symbols, *a, b, c, d, e, f, g,* and *h*, with the following probabilities: $P_a = {}^1/_4$, $P_b = {}^1/_4$, $P_c = {}^1/_8$, $P_d = {}^1/_8$, $P_e = {}^1/_{16}$, $P_f = {}^1/_{16}$, $P_g = {}^1/_{16}$, $P_h = {}^i/_{16}$.

Step 1 creates list 1 and step 2 creates list 2 in Table 1. Step 3 divides the [0-1) interval into 16 equal segments (S0 to S15) and assigns segments S0, S3, S5, S6, S9, S10, S12, and S15 to symbols *a, b, c, d, e, f, g,* and *h* respectively. Step 4 divides each segment into 8 subintervals, creating 8 code points. The first four bits for each code point are marked *x* (*x* could be 0 or 1 according to the segment). Step 5 assigns code points to symbols based on the segment number and the subinterval within the segment.

		list 1	list 2	segment	segment code points	selected code points
a	$P_a = {}^1/_4$	000	0000	S0	0.xxxx00	0.000000
b	$P_b = {}^1/_4$	001	0011	S3	0.xxxx11	0.001111
c	$P_c = {}^1/_8$	010	0101	S5	0.xxxx010	0.0101010
d	$P_d = {}^1/_8$	011	0110	S6	0.xxxx101	0.0110101
e	$P_e = {}^1/_{16}$	100	1001	S9	0.xxxx0110	0.10010110
f	$P_f = {}^1/_{16}$	101	1010	S10	0.xxxx0111	0.10100111
g	$P_g = {}^1/_{16}$	110	1100	S12	0.xxxx1000	0.11001000
h	$P_h = {}^1/_{16}$	111	1111	S15	0.xxxx1001	0.11111001

Table 1: Applying the algorithm to obtain code points with a minimum Hamming distance of 3.

Suppose the string *"bcafadhcbabebagd"* is the source string to be encoded (which corresponds to the source statistics), the encoder for the proposed algorithm will produce the following string:
001111010101000000001010011100000001101011111100101010100011110000000001111100101100011110000000110010000110101
A BSC simulator that introduces errors at a rate of 0.1 resulted in the following string:
$\underline{1}01111010101000\underline{11}001010\underline{10}110000000\underline{0}10101111110010101010001 1\underline{0}1\underline{1}00\underline{1}00001111100101100011110000000 1100\underline{0}000011010\underline{0}$
where the underlined bits are in error. For the original sequence to be constructed, five times the decoder has to consider a single error in the set of bits needed to decode the next symbol; two times the decoder has to consider 2 adjacent errors; and one time 2 non-adjacent errors. Eight times the decoder has to consider no errors. The decoding algorithm maximizes error correction capability by searching for the most likely decoded sequence among all possible sequences based on the following rules:

1- If the symbol to be decoded could be either of two symbols, two separate sequences are considered.
2- If two sequences merge, a MAP decision rejects one and accepts the other.
3- If three errors exist within the set of bits needed to decode the next symbol, the sequence is rejected.
4- If three consecutive double error patterns are detected, the sequence is rejected.
5- A sequence that ends with some residual bits is rejected.

Rule 1 can be shown at the beginning of the string. The first character to be decoded is either *b* (with one bit in error) or *h* (with two bits in error). Any other symbol has $d_{min} > 2$ between that symbol and the received sequence and is not considered. Two separate sequences are considered at this point: one starts with *b* and the other starts with *h*. Notice that the one starting with *b* will start decoding the next symbol from bit number 7 (indexing at the 7th bit), while the one starting with *h* will start decoding the next symbol from bit number 9 (indexing at the 9th bit). Rule 2 applies MAP decisions to eliminate sequences when two or more sequences index at the same bit. This is done based on knowledge of the source statistics and the probability of error. Rule 3 eliminates a sequence if d_{min} between the received string (starting at the indexing bit for this sequence) and each of the eight symbols is greater than two. The string *"hcg"*, which indexes at bit 23, is eliminated based on rule 3. Rule 4 says that if three consecutive double error patterns are decoded, it is enough to alert the decoder to consider the sequence corrupted and to reject it. Sequence "*bcadg*" is eliminated with rule 4. Rule 5 eliminates sequences at the end, since surviving corrupted sequences will not use all the bits in the received data. A total of 31 sequences are considered during decoding. At any given time, no more than 6 sequences survive. Rules 2, 3, and 4 eliminate sequences as more symbols are decoded, while rule 1 creates new sequences. After decoding the twelfth symbol, all the surviving sequences have the same five symbols at the beginning (all split from the correct sequence at a certain point), which means that the decoder can release symbols without having to wait until the end. Only three sequences survive to the end, but because two of them have some residual bits, they are rejected based on rule 5. This maximum likelihood decoder selected *"bcafadhcbabebagd"* as the decoded sequence.

CONCLUSION

The algorithm gives an average of 6.75 bits/symbol. Had conventional arithmetic coding been used, the encoded string would have an average of 2.75 bits/symbol. The algorithm costs a code rate of 11/27. In addition to its error correction capability, the algorithm can provide a high error detection capability. In Table 1, if the source probabilities were equal, the algorithm would have been equivalent to applying the (7,3) block code over the [0-1) interval. For the same errors introduced in the example above, the (7,3) block code would have failed to correct non-adjacent double error patterns, resulting in a block coding error. Notice that applying the algorithm with $d_{min} \geq 3$ allowed the decoder to consider double error patterns. For channels with a lower error rate, making $d_{min} \geq 2$ would allow the design of a decoder that considers single error patterns.

REFERENCES

[1] C. Boyd et al., "Integrating Error Detection into Arithmetic Coding," *IEEE Trans. on Communic.*, vol. 45, no. 1, January 1997, pp. 1-3.

[2] G.F. Elmasry, "An Arithmetic Coding Algorithm With Embeded Channel Coding," *Electronics Ltrs.* vol. 33, issue 20, Sep. 25, 1997, pp. 1687-1688.

[3] G. Langdon, Jr., "An Introduction to Arithmetic Coding," *IBM Jour. of Res.and Dev.*, March 1984, vol. 28, no. 2, pp. 135-149.

Coefficient Rate in Transform Coding*

Wenye Yang
Dept. of Electrical Engineering
Texas A&M University
College Station, TX 77840-3128
wenye@ee.tamu.edu

Jerry D. Gibson
Dept. of Electrical Engineering
School of Engineering and Applied Science
Southern Methodist University
Dallas, TX 75275-0338
gibson@seas.smu.edu

Abstract

In transform coding, the energy of transform coefficients varies from sample function to sample function, and at low rate, many coefficients will be quantized to zero. By eliminating small energy coefficients, we can reduce the number of coefficients to be coded, but side information as to which coefficients are coded is needed for decoders to decode the coded sequence. In this paper, we investigate a concept defined by Campbell, the *coefficient rate* of random processes and develop a theory that can be used in transform coding to determine the number of coefficients to be coded and to estimate the code rate used for encoding the side information.

1 Introduction

In transform coding, a source is usually modeled as a stationary random process. The sequence generated by the source is cut into pieces and each piece is treated as a sample function of the process. From sample function to sample function, the energy of a transform coefficient varies over a wide range. After quantization, many coefficients are quantized to zero but the distribution of zeroes among components varies from sample function to sample function. For an encoder, it would be desirable if it could avoid coding these zeroes, but for decoders to decode the code sequence, information on which coefficients are zero and which are not is needed. Such side information is generally hard to characterize, and therefore it is difficult to design a rule for the coder to efficiently send such information.

Among several existing adaptive schemes, the use of *spectral entropy* as an indicator of the number of coefficients to be coded can be found in different applications[1]-[3], but so far, no theory has been offered as to why spectral entropy should be a good indicator.

In this paper, we investigate a concept defined by Campbell[4], the *coefficient rate* of random processes, which is closely related to the spectral entropy, and develop a theory to determine the number of coefficients to be coded in a random source and to estimate the code rate for side information.

2 Coefficient Rate of Random Processes

2.1 Treating N sample functions together

The idea behind Campbell's approach is intuitively quite straightforward. When a sequence of samples of a random variable is to be coded, it is generally true that concatenating samples and coding them together can achieve a lower code rate. When the

---

*This research is supported in part by NSF Grant No. NCR-9627610

number of samples concatenated gets very large, the code rate can approach its lower bound, the *entropy* of the random variable, as close as possible. The same idea is used by Campbell when a number of sample functions of a random process are considered.

Consider a zero-mean stationary random process $X(t)$, using the Karhunen-Loeve expansion (K-L expansion), in time duration $[0, T]$, the process can be decomposed as

$$X(t) = \sum_{i=1}^{\infty} C_i \phi_i(t) \tag{1}$$

where C_i's are random variables with zero mean and $\mathrm{E}[C_i^2] = \lambda_i$ and $\phi_i(t)$ are eigenfunctions of the expansion. Let $x_1(t_1), x_2(t_2), \ldots, x_N(t_N), 0 < t_j < T, j = 0, 1, \ldots, N$ be N independent sample functions of the process, and assuming there are only finite components in the K-L expansion, these sample functions can be expressed as

$$x_j(t_j) = \sum_{i=1}^{M} c_{ji} \phi_i(t_j) \qquad j = 1, 2, \ldots, N \tag{2}$$

where $c_{ji}, j = 1, 2, \ldots, N$ are N samples of random variable C_i .

Let $y(t_1, t_2, ..., t_n) = x_1(t_1)x_2(t_2) \cdots x_N(t_N)$, the product of sample functions, and since these sample functions are independent, $y(t_1, t_2, ..., t_N)$ can be expressed in terms of products of eigenfunctions of the K-L expansion,

$$\begin{aligned} y(t_1, t_2, ..., t_n) &= \sum_{i_1, i_2, ..., i_N = 1}^{M} c_{i_1} c_{i_2} \cdots c_{i_N} \phi_{i_1}(t_1) \phi_{i_2}(t_2) \cdots \phi i_N(t_N) \\ &= \sum_{k=1}^{M^N} c^{(k)} \phi^{(k)}(t_1, t_2, \ldots, t_N) \end{aligned} \tag{3}$$

where $c^{(k)}$ is the product of c_{i_j} and is arranged in decreasing order of its variance, and $\phi^{(k)}(t_1, t_2, \ldots, t_N)$ is also reordered accordingly.

Campbell tried to approximate $y(t_1, t_2, ..., t_N)$ by choosing $\mu (< M^N)$ $c^{(k)}$'s with largest variances such that the average energy loss of the approximation is small. To find μ, let

$$y_\mu(t_1, t_2, ..., t_N) = \sum_{k=1}^{\mu} c^{(k)} \phi^{(k)}(t_1, t_2, ..., t_N) \tag{4}$$

such that

$$\mathrm{E}[y(t_1, t_2, ..., t_N) - y_\mu(t_1, t_2, ..., t_N)]^2 < \epsilon. \tag{5}$$

Campbell used an AEP-like argument [4] to estimate μ, which satisfies

$$\lim_{T \to \infty} \lim_{N \to \infty} \log \frac{\mu^{1/N}}{T} = -\int_{-\infty}^{\infty} S(f) \log S(f) df \tag{6}$$

where $S(f)$ is the power spectral density of the random process. He then argued that $[\mu^{1/N}]/T$ is the number of coefficients per dimension per unit time, so he defined the *minimum coefficient rate* of a random process as

$$Q = \exp\Big[-\int_{-\infty}^{\infty} S(f) \mathrm{log} S(f) df \Big] \tag{7}$$

By using the AEP more directly, we can get a similar result [5] that relates Q with signal bandwidth. The equivalent bandwidth W_e, as defined in [5], and the coefficient rate Q have a simple relation: $Q = 2W_e$.

2.2 Dominant terms in the expansion

We can explore this result even further to asymptotically estimate the variance of the surviving terms after truncation in (4) as $N \to \infty$ and to calculate the number of these terms. Since all of the sample functions are independent, for any random variable $C^{(k)} = C_{i_1} C_{i_2} \cdots C_{i_N}$, $\mathrm{E}[(C^{(k)})^2] = \mathrm{E}[C_{i_1}^2]\mathrm{E}[C_{i_2}^2] \cdots \mathrm{E}[C_{i_N}^2] = \lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_N}$. So

$$\begin{aligned}\mathrm{E}[y(t_1,t_2,\ldots,t_N)^2] &= \sum_{k=1}^{M^N} \mathrm{E}[(C^{(k)})^2] = \sum_{M^N terms} \lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_N} \\ &= \prod_{j=1}^{N}\sum_{i=1}^{M} \lambda_i = T^N. \end{aligned} \tag{8}$$

There are a total of M^N terms in $\mathrm{E}[y(t_1,t_2,\ldots,t_N)^2]$, but some of them are very small compared to others. As $N \to \infty$, those large energy terms will dominant the sum and small energy terms can be thrown away without affecting the total energy, and hence the number of terms that preserve the energy of $y(t_1,t_2,\ldots,t_N)$ can be reduced.

From (8), we have

$$\begin{aligned}\mathrm{E}[y(t_1,t_2,\ldots,t_N)^2] &= \prod_{j=1}^{N}\sum_{i=1}^{M} \lambda_i = \Big(\sum_{i=1}^{M}\lambda_i\Big)^N \\ &= \sum_{\{n_i, i=1,2,\ldots,M: \sum_{i=1}^{M} n_i = N\}} \frac{N!}{n_1! n_2! \cdots n_M!}\lambda_1^{n_1}\lambda_2^{n_2}\cdots\lambda_M^{n_M}. \end{aligned} \tag{9}$$

We want to find the largest term of $\frac{N!}{n_1!n_2!\cdots n_M!}\lambda_1^{n_1}\lambda_2^{n_2}\cdots\lambda_M^{n_M}$ subject to constraint that $\sum_{i=1}^{M} n_i = N$. As $N \to \infty$, this term will grow much faster than any other terms, and it will dominate the total energy of $y(t_1,t_2,\ldots,t_N)$. So the problem becomes finding the maximum of the functional

$$I = \log\Big\{\frac{N!}{n_1!n_2!\cdots n_M!}\lambda_1^{n_1}\lambda_2^{n_2}\cdots\lambda_M^{n_M}\Big\} + \alpha\sum_{i=1}^{M} n_i. \tag{10}$$

Using the approximation $\log N! = N\log N - N$ for large N, we have

$$\begin{aligned} I &= \log N! - \sum_i \log n_i! + \sum_i n_i \log\lambda_i + \alpha\sum_i n_i \\ &= N\log N - \sum_i n_i\log n_i + \sum_i n_i\log\lambda_i + \alpha\sum_i n_i. \end{aligned} \tag{11}$$

Taking partial derivatives of I with respect to n_i and setting them to zero, we have

$$\frac{\partial I}{\partial n_i} = -\log n_i - 1 + \log\lambda_i + \alpha = 0,$$

so, $n_i = \mathrm{e}^{\alpha-1}\lambda_i$. By using the constraint $\sum_i n_i = N$, we obtain $\alpha - 1 = \log(N/T)$. So,

$$n_i = \frac{\lambda_i}{T}N \qquad i = 1,2,\ldots,M. \tag{12}$$

That is, the number of λ_i in $\lambda_1^{n_1}\lambda_2^{n_2}\cdots\lambda_M^{n_M}$ is proportional to λ_i.

We can find the average energy and the number of these dominant terms. We get the energy

$$\sigma^2 = \lambda_1^{n_1}\lambda_2^{n_2}\cdots\lambda_M^{n_M} = T^N\mathrm{e}^{-NH(S)}, \tag{13}$$

and the number of such terms is

$$\mu = \frac{N!}{n_1! n_2! \cdots n_M!} = e^{NH(S)} \tag{14}$$

where $H(S) = -\sum_i (\lambda_i/T)\log(\lambda_i/T)$, the spectral entropy in discrete form. This agrees with Campbell's result. We also have the asymptotic relation $T^N = \mu\sigma^2$, and we can see that when $N \to \infty$, we can use these μ equal energy terms to approximate $y(t_1, t_2, \ldots, t_N)$, and energy loss introduced by truncation is negligible.

3 Coefficient Rate in Source Coding

3.1 Coding coefficients according to their significance

The above result says that we can use fewer coefficients to approximate a signal without losing significant energy. This suggests a way to reduce the number of coefficients to be coded in transform coding problems.

Consider N sample functions of a random source, each sample function can be represented by a set of coefficients after the K-L expansion. Here we use *component* to refer to the random variable after the K-L expansion and *coefficient* refers to a sample of the random variables, or components. To reduce the number of coefficients to be coded, we have to design a rule to determine which coefficients are to be coded.

Based on the rate distortion theory, the general way is to use optimal bit allocation to decide which components are to be coded. If a component has an average energy higher than the quantization error, all coefficients of this component will be coded. So for an individual coefficient, whether it is coded or not depends only on the average energy of the component to which it belongs.

We design a new rule to determine if a coefficient is coded or not depending on not only the average energy of the component but also the energy of that individual coefficient. To reduce the number of coefficients to be coded, we do not simply drop all coefficients of a low energy component, instead, we determine the number of coefficients to be coded according to the significance of the component. The distribution of components in those dominant terms shows that component energy acts as a measure of contribution of the component to the total energy. Therefore, component energy can be used as a measure of significance of components. If a component is significant, more coefficients of that component will be encoded and the number of coefficients to be coded is proportional to the average energy of the component. By this means, we eliminate more coefficients of small energy components, and the remaining coefficients after the reduction have a distribution similar to those dominant terms.

For individual coefficients, whether it is coded or not depends on two factors: the average energy of the components and its own energy. If the average energy of a component is high, coefficients belong to the component are more likely to be coded. Since only a fraction of coefficients of the component will be coded, those with highest individual energies are chosen.

A simple example shows what this means to coding. Let X_i be a component of the random process and assume it is Gaussian with zero mean and variance λ_i. Then in bits per coefficient, the code rate for X_i is

$$R_i = \frac{1}{2}\log\frac{\lambda_i}{D_i} \tag{15}$$

where D_i is the distortion level. Suppose we have N samples of X_i but many of them can be quantized to zero, then we can reduce the number of samples to be coded to n_i, although it is not necessary that we eliminate all zero samples.

Now with only n_i samples but almost the same total energy, we have a new average energy among these n_i samples, $\lambda_i' = N\lambda_i/n_i$. To estimate the code rate, still using the Gaussian model, and at the same distortion level,

$$R_i' = \frac{1}{2}\log\frac{\lambda_i'}{D_i} = \frac{1}{2}\log\frac{\lambda_i}{D_i} + \frac{1}{2}\log\frac{N}{n_i}. \tag{16}$$

The first term is just the original code rate and the second term depends on n_i, the number of samples to be coded. Since we select the number of samples of a component in proportion to its energy, this term carries information as to which component it represents.

Similarly, for a random process, the total code rate consists of two parts: one is that from the first term, which carries the actual information of coefficient values; the other is from the second term, which carries what we refer to as *side information* and has direct relation with the spectral entropy.

3.2 Encoding the significance map

Using component energy as a significance measure, we can determine how many coefficients should be coded for each component. Then, whether an individual coefficient is coded depends on its own energy relative to other coefficients from the same component. Coefficients with higher energy will be chosen first.

Due to the randomness of the source, the distribution of coded coefficients will be different from sample function to sample function, therefore, information of which coefficients are coded should also be encoded. A significance map is a representation of those transform coefficients with sufficient energy that they must be coded and transmitted to achieve acceptable reconstructed signal quality. For each sample function, a significance map can be constructed to indicate which coefficients in the sample function are to be coded. Then we can just send the index of coefficients in the significance map, since we already know the distribution of these indices, we can use spectral entropy to estimate the code rate used for the significance map .

3.3 The information rate of coefficient values

We now consider how to encode those selected coefficients. Let X^L represent a sequence of length L in which coefficients are chosen from samples of M random variables $X_1, X_2, \ldots, X_M$ and the probability of choosing X_i is proportional to its variance λ_i. Then for L sufficiently large, the number of X_i's in the sequence should be close to $p_i L$.

Let $\hat{X}^L$ represent the coded sequence of X^L, the rate distortion function is then defined on a per coefficient basis as

$$R(D) = \min_{\mathrm{E}d(X^L,\hat{X}^L)\le D} \frac{1}{L} I(X^L, \hat{X}^L) \tag{17}$$

where $d(X^L, \hat{X}^L)$ is some distortion measure. We have

$$I(X^L, \hat{X}^L) \;=\; h(X^L) - h(X^L|\hat{X}^L) = \sum_{i=1}^{M} p_i L h(X_i) - \sum_{j=1}^{L} h(X_j|\hat{X}^L)$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{M} p_i L h(X_i) - \sum_{j=1}^{L} h(X_j|\hat{X}_j) && (18)\\
&= \sum_{i=1}^{M} p_i L[h(X_i) - h(X_i|\hat{X}_i)] \\
&\geq \sum_{i=1}^{M} p_i L R(D_i). && (19)
\end{aligned}
$$

If we choose $f(x^L, \hat{x}^L) = \prod_{j=1}^{L} f(x_j, \hat{x}_j)$, i.e., samples are independent, then equality holds in (18).

For a Gaussian process, after the K-L expansion, components of the expansion are independent Gaussian random variables, thus we can choose $\hat{X}_i$ as a random variable $\sim N(0, \lambda_i - D_i)$ so that equality holds in (19).

Then for a Gaussian random process, the rate distortion problem becomes optimizing

$$R(D) = \min_{\sum_{i=1}^{M} p_i D_i \leq D} \Big\{ \sum_{i=1}^{M} p_i \Big(\frac{1}{2} \log \frac{\lambda_i}{D_i}\Big) \Big\}. \tag{20}$$

Using Lagrange multipliers, we have the functional

$$J(D) = \sum_{i=1}^{M} p_i \Big(\frac{1}{2} \log \frac{\lambda_i}{D_i}\Big) + \alpha \sum_{i=1}^{M} p_i D_i. \tag{21}$$

Taking partial derivatives with respect to D_i and setting them to zero, we get

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2} \frac{p_i}{D_i} + \alpha p_i = 0 \tag{22}$$

so $D_i = 1/2\alpha$ and using $\sum_{i=1}^{M} p_i D_i = D$, we get $D_i = D$, assuming all $\lambda_i \geq D_i$. Then we have the rate distortion function

$$R(D) = \sum_{i=1}^{M} p_i \Big(\frac{1}{2} \log \frac{\lambda_i}{D}\Big). \tag{23}$$

If we let D get large enough that some λ_i may be less than D. Then the rate distortion function is

$$R(D) = \sum_{i=1}^{M} p_i \Big(\frac{1}{2} \log \frac{\lambda_i}{D_i}\Big) \tag{24}$$

where

$$D_i = \begin{cases} D & \lambda_i > D \\ \lambda_i & \lambda_i < D \end{cases}$$

This is very similar to the classical rate distortion theory for one sample function. But by considering N sample functions together, samples of high energy components are more likely to be encoded and the total number of coefficients being coded is reduced.

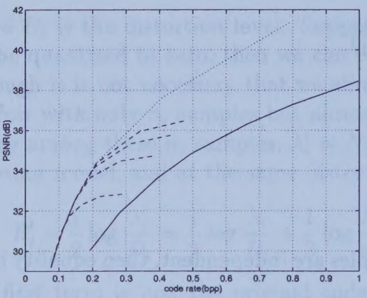


Figure 1: Effects of coefficient reduction

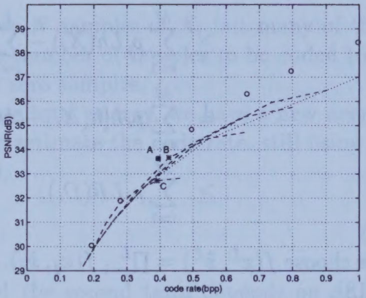


Figure 2: Coder performances

4 Experiment Results

4.1 Reducing the number of coefficients

In our experiment, we used the 512 × 512 *lena* image as the source and 8 × 8 DCT as the transform, so the number of components is 64 and there are 4096 blocks, which we treated as independent sample functions of the image, although coefficients in these blocks are, in general, highly correlated.

For comparison, we also implemented a DCT-based image coder using optimal bit allocation. To encode coefficients allocated nonzero bit rate, uniform quantizers are used followed by entropy coders, which in this case, are adaptive arithmetic coders. In terms of SNR, this image coder should be considered optimal.

First we would like to see how much difference it makes if we avoid sending coefficients that are quantized to zero and the effects of reducing the number of coefficients to be coded. Fig. 1 shows the difference between coding with and without zeroes and the energy loss after reducing the number of coefficients. The dotted curve shows the relation of PSNR vs. code rate if all the coefficients quantized to zero are not encoded, and the solid curve is the result of the image coder using optimal bit allocation. We can see that there is a nearly $4dB$ difference between the two curves, or just half of the code rate is actually used for nonzero coefficients.

To reduce the number of coded coefficients n_i for each component, we first estimated the average energy of each component. This was done by calculating the variance of the DCT coefficients among all 4096 blocks.

Since we only want to keep the distribution of coefficients close to the estimated spectrum of the source, we still can change n_i, the number of coded coefficients. In our experiment, we used a factor f to control n_i. When $f = 1$, if the jth component has the largest average energy, we set $n_j = 4096$. If we want to increase the number of coefficients, we increase f so that all n_i change accordingly. However, it can be seen that if $f > 1$, it is possible that for high energy coefficients, n_i may be larger than 4096, in that case, we just set them to be 4096.

The dashed curves in Fig. 1 shows the effects of reducing the number of coefficients. The lowest curve is for $f = 1$, the others are for $f = 2, 3, 4$ respectively and the larger the f, the closer the curve is to the dotted one. By fixing n_i before quantization, we are actually setting a threshold for the coefficients and when the quantization level is low, this threshold may be higher than the quantization level. This is why energy loss can be seen in Fig. 1 as we lower the quantization level, since some coefficients are dropped by thresholding even if they may not be quantized to zero. But at low rate, when the quantization level is high, all the dashed curves converge to the dotted one.

4.2 Image coder using the significance map method

Since coefficients being coded in each sample function are different from other sample functions, for a real coder, side information must also be coded so that the decoder knows which coefficient is actually coded.

To encode the significance map, the indices of coefficients in the significance map for each block are encoded using an adaptive arithmetic coder. Then, the coefficient values are encoded using a uniform quantizer followed by an adaptive arithmetic coder, as was done by the image coder using optimal bit allocation.

Fig. 2 compares the performance of the two image coders. The big dots in the figure are results from the image coder with optimal bit allocation, and the dashed curves represent the results by coders using the significance map method. Still, the lowest curve is for $f = 1$. At low rate, performance of the coder using the significance map is nearly as good as the coder using optimal bit allocation. But when the quantization level gets lower, energy loss due to thresholding shows up.

At high rate, when the number of coefficients is increased to a point where the distribution among coefficients is no longer close to the spectrum of the source, encoding the significance map as describe here is not efficient. One of the reasons is that the method used here to encode the significance map treats indices as they are in random order, which at low rate is nearly true. But within a block, indices are ordered, and this order also provides valuable information. At high rate, such information has to be considered in order to efficiently encode the significance map.

4.3 Comparison of images

It is also interesting to directly compare the coded images. We take three images, which have very close code rates and PSNR, corresponding to points A, B, C in Fig. 2 to compare the subjective effects of the coders. Although the image corresponding to point C has the lowest PSNR among the three, it has less blocking effects and hence looks smoother in large slowly changing areas.

Fig. 3 shows the original image and the three coded images. To see the differences of blocking effects more clearly, only edges in these images are shown here. Here edges mean differences from neighboring pixels. We can see that images (c) and (d), corresponding to points B and A in Fig. 2 respectively, have more clear (white) areas than in image (b). Since there is no change in the clear area, in actual coded images, transitions in these areas are totally lost, as a result, we feel the blocking effects in these areas. On the other hand, image (b) shows more changes in the background area, which makes image (b) look smoother perceptually.

This is because in image (b), the number of coefficients being coded is less, so the coded coefficients enjoy lower quantization level, therefore they are encoded with higher fidelity. At low rate, encoding the low frequency components with high fidelity often results in a better image subjectively. But the cost is a higher thresholding for high frequency components.

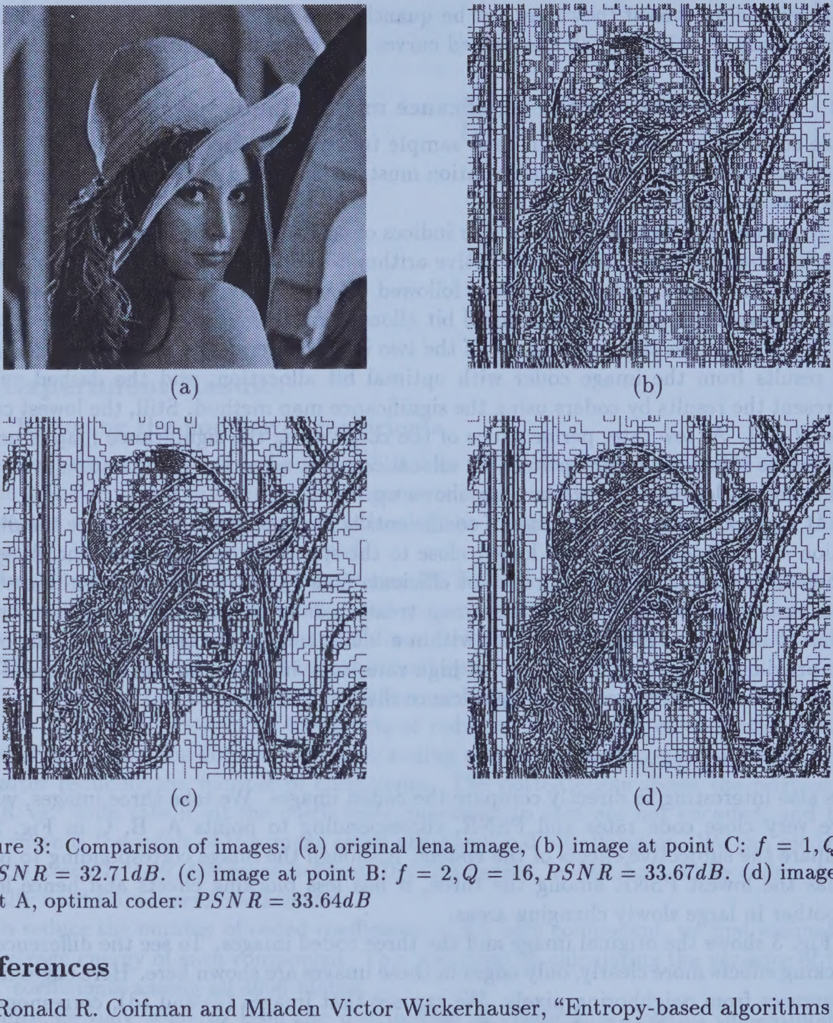


Figure 3: Comparison of images: (a) original lena image, (b) image at point C: $f = 1, Q = 8, PSNR = 32.71dB$. (c) image at point B: $f = 2, Q = 16, PSNR = 33.67dB$. (d) image at point A, optimal coder: $PSNR = 33.64dB$

References

[1] Ronald R. Coifman and Mladen Victor Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 713-718, March 1992.

[2] R. Mester and U. Franke, "Spectral entropy-activity classification in adaptive transform coding," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 913-917, June 1992.

[3] Stan McClellan and Jerry D. Gibson, "Variable-Rate CELP Based on Subband Flatness," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 120-130, March 1997.

[4] L. Campbell, "Minimum coefficient rate for stationary random processes," *Information and Control*, vol. 3, no. 4, Dec. 1960.

[5] Wenye Yang and Jerry D. Gibson, "Spectral entropy, equivalent bandwidth and minimum coefficient rate," in *Proc. of ISIT'97*, Ulm, Germany, June 29 - July 4, 1997.

[6] N. S. Jayant and Peter Noll, *Digital Coding of Waveforms*, Prentice-Hall, 1984.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

Modeling Video Traffic Using Wavelets

Sheng Ma and Chuanyi Ji
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180
shengm@ecse.rpi.edu and chuanyi@ecse.rpi.edu

Abstract

In this work, we establish a wavelet model on video traffic. Different from the existing methods which model the video traffic in the time domain, we model the wavelet coefficients in the wavelet domain. The strength of the wavelet model includes (1) a unified approach to model both the long-range and the short-range dependence in the video traffic simultaneously, (2) the ability to reduce the temporal dependence so significantly that the wavelet coefficients can be modeled by either independent or Markov models in the wavelet domain, (3) a computationally efficient method on developing the model and generating high quality video traffic, and (4) feasibility of performance analysis using the model.

1 Introduction

One of the significant statistical properties found recently on VBR video traffic is the co-existence of the long-range (LRD) and the short-range (SRD) dependence (see for example [2][3] and references herein). Intuitively, this property results from scene changes, and suggests a complex behavior of video traffic in the time domain[5]. This complex temporal behavior makes accurate modeling of video traffic a challenging task. The goal of this work is to develop a unified and computationally efficient method to model both the long-range and the short-range dependence in real video sources.

Ideally, a good traffic model needs to be *(a)* accurate enough to characterize pertinent statistical properties in the traffic, *(b)* computationally efficient, and *(c)* feasible for the analysis needed for network design. The existing models developed to capture both the long-range and the short-range dependence include ARIMA models[2], a model based on Importance Sampling[3], Transform-Expand-Sample (TES) model[7] and scene-based models[5]. All these methods model both LRD and SRD in the time domain. The scene-based modeling[5] provides a physically interpretable model feasible for analysis but difficult to be made very accurate. TES method is reasonably fast but too complex for the analysis. The rest of the methods suffer from the computational complexity too high to be used to generate a large volume of synthesized video traffic.

To circumvent these problems, we will model the video traffic in the wavelet domain instead of in the time domain. Motivated by the previous work on wavelet representations of (the LRD alone) Fractional Gaussian Noise (FGN) process (see [8] and references herein), we will show in this paper simple wavelet models can simultaneously capture the short-range and the long-rage dependence through modeling two video traces. Since the (deterministic) self-similar structure of wavelets provides a natural match to the (statistical) self-similarity of the long-range dependence, wavelet coefficients become short-range dependent[8]. This suggests that modeling can be a lot easier if done in the wavelet domain. As the result, we will show that models based on independent wavelet coefficients will be reasonably accurate for matching the auto-correlation function and the buffer loss probability. More complex models such as the first order Markov models for modeling the dependence of wavelet coefficients between time-scales will only improve the performance marginally. Furthermore, since the computation needed for wavelet transforms and inverse transforms is in the order of $O(N)$ for a video trace of length N, our approach will be able to attain the lowest computational complexity to generate wavelet models. Through our theoretical analysis on the buffer loss rate, we will also demonstrate the feasibility of using wavelet models for theoretical analysis.

1.1 Wavelet Transforms

In $L^2(R)$ space, discrete wavelets $\phi_j^m(t)$'s are ortho-normal basis which can be represented as $\phi_j^m(t) = 2^{-j/2}\phi(2^{-j}t - m)$, for $t \in [0, 2^K - 1]$ with $K \geq 1$ being an integer. $\phi(t)$ is the so-called mother wavelet. $1 \leq j \leq K$ and $0 \leq m \leq 2^{K-j} - 1$ represent the time-scale and the time-shift, respectively. Since wavelets are the dilation and shift of a mother wavelet, they possess a self-similar structure at different time scales. For simplicity, the mother wavelet in this work is chosen to be the Haar wavelet, where $\phi(t)$ is 1 for $0 \leq t < 1/2$, -1 for $1/2 \leq t < 1$ and 0 otherwise.

Let d_j^m's be wavelet coefficients of a discrete-time process $x(t)$ ($t \in [0, 2^K - 1]$). Then d_j^m can be obtained through the wavelet transform $d_j^m = \sum_{t=0}^{2^K-1} x(t)\phi_j^m(t)$. $x(t)$ can be represented through the inverse wavelet transform $x(t) = \sum_{j=1}^{K} \sum_{m=0}^{2^{K-j}-1} d_j^m \phi_j^m(t) + \phi_0$, where ϕ_0 is equal to the average of $x(t)$.

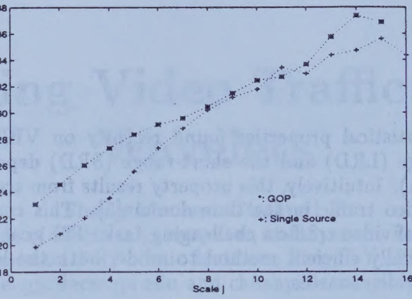


Figure 1: Log 2 of Variance of d_j versus the time scale j

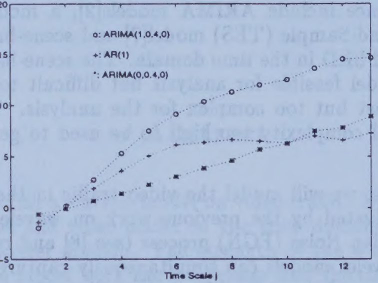


Figure 2: Log 2 of Variance of d_j versus the time scale j

2 Wavelet Modeling of Video Traffic

2.1 The Video Sources

Two video sources are used to test our wavelet models: (1) "Star Wars"[2], where each frame is encoded by JPEG-like encoder, and (2) MPEG coded videos at Group of Pictures (GOP) level[5][10] called "MPEG GOP" in the rest of the paper. The modeling is done at either the frame level or the GOP level.

2.2 The Variances and Auto-correlation of Wavelet Coefficients

As the first step to understand how wavelets capture the LRD and SRD, we plot in Figure (1) the variance of the wavelet coefficients d_j^m's at different time scales for both sources. To understand what the curves mean, we also plot in Figure (2) the variances of wavelet coefficients for three well-known processes: ARIMA$(0, 0.4, 0)$, ARIMA$(1, 0.4, 0)$, and AR(1). ARIMA$(0, 0.4, 0)$ is a long-range dependent process with Hurst parameter $H = 0.9$. AR(1) is a short-range dependent process, and ARIMA$(1, 0.4, 0)$ is a mixture of the long-range and the short-range dependent process.

As observed, for ARIMA$(0, 0.4, 0)$ process (LRD alone), the variance increases with j exponentially for all j. For AR(1) (SRD alone), the variance increases at an even faster rate than that of ARIMA$(0, 0.4, 0)$ when j is small but saturates when j is large. For ARIMA$(1, 0.4, 0)$, the variance shows the mixed properties from both AR(1) and ARIMA$(0, 0.4, 0)$. The variance of the video sources behaves similarly to that of ARIMA$(1, 0.4, 0)$, and thus demonstrate the co-existence of the SRD and LRD in the video sources in the wavelet domain. This also indicates through the second moment that wavelets are capable of capturing both the long-range and the short-range dependence[1].

2.3 Correlation Structure of Wavelet Coefficients

One of the key advantages for modeling in the wavelet domain is the ability of wavelet transforms to significantly reduce (the long-range) dependencies. In particular, analytical results on the correlation structure of the wavelet coefficients have been obtained by [11][4] for a (long-range-dependent) FGN process, and can be summarized as follows.

Theorem 1 **(Kaplan and Juo[4])** *Let $x(t)$ be a FGN process with Hurst parameter H $(0.5 < H < 1)$. Let d_j^m's be the (Haar) wavelet coefficients of $x(t)$. Then*
(1) for a given time scale j, d_j^m's are i.i.d. Gaussian random variables with zero mean and variance $2^{j(2H-1)}(2^{2(1-H)}-1)\sigma^2$, where σ is the variance of $x(t)$.
(2) for $(m_1+1)2^{j_1} - m_2 2^{j_2}$ large, where j_1, j_2, m_1 and m_2 are the dilation and the translation indices of two different wavelet coefficients respectively, the correlation between two wavelet coefficients is

$$E(d_{j_1}^{m_1} d_{j_2}^{m_2}) \sim O(|2^{j_1}m_1 - 2^{j_2}m_2|^{-2(2-H)}). \tag{1}$$

We recall that the corresponding temporal auto-correlation of Fractional Gaussian Noise decays at a rate $O(|k|^{-2(1-H)})$, where k is the lag between two samples and $0.5 < H < 1$. This rate leads to a divergent summation of the auto-correlations. Theorem 1 indicates that the wavelet transformation has changed the long-range-dependence in the time domain so significantly that the summation of the (auto-)correlation of wavelet coefficients converges to a constant.

For a shot-range dependent process, we derive analytical results on the correlation structure in the wavelet domain in a theorem given below.

Theorem 2 *Let $x(t)$ be a zero-mean wide-sense-stationary Gaussian process with the auto-correlation $r(k)$ of $x(t)$ satisfying $r(k) = \sigma^2\rho^{|k|}$, where $|\rho| < 1$, k is an integer and σ^2 is the variance of $x(t)$. Let d_j^m's be the (Haar) wavelet coefficients of $x(t)$. Then*
(1) for a given time scale j, d_j^m's are i.i.d. Gaussian random variables with zero mean and variance $\sigma^2(1 + \frac{2\rho}{1-\rho} - \frac{3\rho}{(1-\rho)^2 2^j - 1}) + O(\rho^{2^j-1})$.
(2) for $m_1 2^{j_1} - (m_2+1)2^{j_2} > 0$,

$$E(d_{j_1}^{m_1} d_{j_2}^{m_2}) = 2^{\frac{-j_1-j_2}{2}} \rho^{m_1 2^{j_1} - (m_2-1)2^{j_2}} (1-\rho^{2^{j_2-1}})^2 (1-\rho^{2^{j_1-1}})^2 \frac{\rho}{(1-\rho)^2}\sigma^2. \tag{2}$$

Theorem 2 shows that the correlation of wavelet coefficients decays exponentially at an even faster rate than the corresponding auto-correlation in the time domain.

Combining Theorems 1 and 2, we can see that the wavelet transforms significantly reduce the temporal dependence so that simple models with "short-range" dependence may be sufficient in the wavelet domain to model the complicated mixture of the short- and the long- range dependence in the time domain. This provides a strong motivation on modeling the traffic in wavelet domain.

---

[1] The analytical expressions of the variances for the long-range and the short-range dependent Gaussian processes can be found in [6].

2.4 Modeling The Correlation Structure

To motivate how to model the correlation in the wavelet domain, we plot the correlation matrix ARIMA(1,0.4,0) in Figure (3), where ARIMA(1,0.4,0)[2] is a Gaussian process with a mixture of the long- and short-range dependence. The diagnal line alone corresponds to independent wavelet coefficients. The four pairs of off-diagnal lines corresponds to correlations among wavelet coefficients at adjacent time-scales. Longer term correlations are invisible in the figure. This indicates that the most significant correlation is due to the correlations between time scales. Since the complicated correlation in the time domain actually concentrates on these types of correlation in the wavelet domain, we can use a parsimonious model in wavelet domain to faithfully represent the original traffic. Several such models have been investigated systematically in [6], and two of which are summarized as follows:

Model 1: An independent model which neglects the dependence among wavelet coefficients. That is, d_j^m's are chosen to be independent Gaussian random variables with a zero mean and a variance, σ_j. The variance is estimated from data for each dilation index j[3].

Model 2: A first-order Markov model which captures the most significant correlation due to the wavelet coefficients between time scales.

Model 3: A third-order Markov model which captures the longer-term (but weaker) correlations among wavelet coefficients.

To evaluate the performance of our models, we plot the sample autocorrelation and buffer response due to both the wavelet model and ARIMA(1,0.4,0) in Figures (4) and (5), respectively. We observe that the simplest wavelet model which neglects the dependence in the wavelet domain performs reasonably well. The models which capture more correlations among wavelets only yield marginally improved performance.

2.5 Marginal Probability Density Functions

What are the marginal distributions of wavelet coefficients for real video traffic? Figure (6) plots the $Q-Q$ plots for the wavelet coefficients of the two sources at $j=3$[4]. The figure shows that the sample marginal density functions of wavelet coefficients for both the "Star Wars" and the MPEG GOP source at the given time scale have a much heavier tail than that of the normal distribution. Therefore, the variance alone is only sufficient when the marginal density function is normal, and in general a marginal density function should be considered as another pertinent statistical property.

2.6 An Algorithm on Generating Wavelet Models for Video Traffic

The algorithm we derive include three main steps: (a) obtain sample variances of wavelet coefficients at each time scale, (b) generate wavelet coefficients independently[5] from the normal marginal density function using the sample mean and variance, and (c) perform a transformation on the wavelet coefficients so that the resulting wavelet coefficients have a marginal density function required by the traffic. The obtained wavelet coefficients form a wavelet model from which synthesized video traffic can be generated. The algorithm can be summarized as follows.

---

[2] Samples of ARIMA(1,0.4,0) are generated by Splus. We choose the AR parameter to be 0.9 and the differential parameter to be 0.4 which corresponds to Hurst parameter 0.9.

[3] It can be shown that d_j^m is stationary in terms of either fixed j or m.

[4] Similar behaviors have been observed at the other time scales. A $Q-Q$ plot is a standard statistical tool to measure the deviation of a marginal density function from a normal density. The $Q-Q$ plots of a process with a normal marginal is a straight line. The deviation from the line indicates the deviation from the normal density. See [2] and references herein for more details.

[5] For simplicity, we only consider the algorithm which generates independent wavelet coefficients. Descriptions on the Markov models used for generating wavelet coefficients can be found in [6].

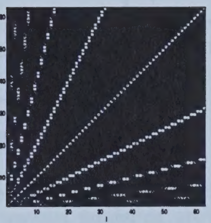


Figure 3: Correlation Matrix of ARIMA(1,0.4,0).A pixel (i, j) in the image represents the correlation between the i-th and the j-th wavelet coefficients. The gray level is proportional to the absolute value of the correlation.

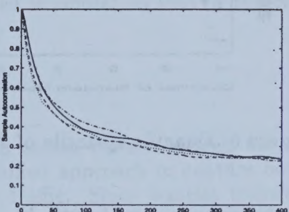


Figure 4: Sample Correlations. "-": ARIMA(1,0.4,0);"-.": Model 3; "- -": model 2; "..": model 1.

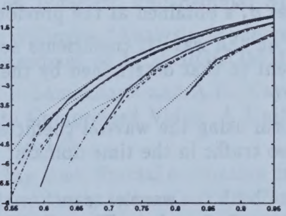


Figure 5: Buffer Response. x-axis: utilization. y-axis: log10(Overflow probability). "-": ARIMA(1,0.4,0);"--": Model 3; "-.": model 2; "..": model 1. The normalized buffer size B/C: 0.1, 1, 10, and 100 from the top down.

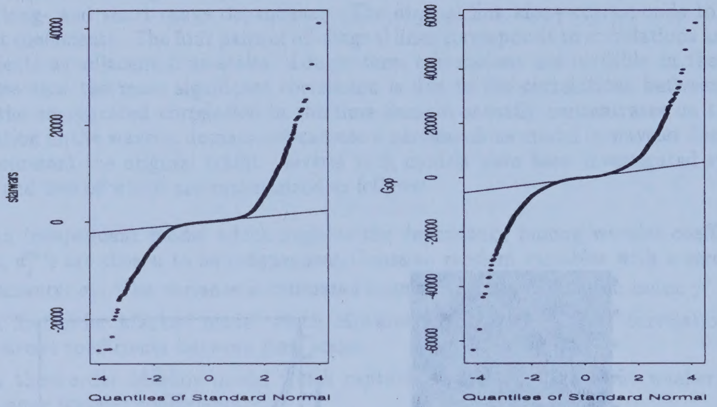


Figure 6: Quantile-Quantile of d_j^m for $j = 3$. Left: Star Wars. Right: GOP.

Let $\hat{x}(t)$ be the video trace of length N.

Algorithm

1. Obtain wavelet coefficients from $\hat{x}(t)$ through the wavelet transform.
2. Compute the sample variance $\hat{\sigma}_j$ of wavelet coefficients at each time scale j using what obtained at the previous step.
3. Generate new wavelet coefficients d_j^m's for all j and m independently through Gaussian distributions with variances $\hat{\sigma}_j$'s obtained at the previous step.
4. Perform a transformation on the wavelet coefficients so that the corresponding marginal density function is consistent to that determined by the video traffic(see [3] for details on the transformation).
5. Do inverse wavelet transform using the wavelet coefficients obtained at the previous step to get the synthesized video traffic in the time domain.

The computational complexity of both the wavelet transform (Step 1) and the inverse transform (Step 5) is $O(N)$. So is for Steps 2, 3 and 4. Then $O(N)$ is the computational cost of the algorithm, which is the lowest attainable for traffic models.

2.7 Experimental Results

Video traces of length 171,000 for "Star Wars" and 66369 for "MPEG GOP" are used to obtain wavelet models. ARIMA models with 45 parameters are also obtained using the same data for comparison. The synthesized video traffic from both models are generated and used to obtain sample auto-correlation functions in the time-domain, and to estimate the buffer loss rate. The results[6] are given in Figure (8). Wavelet models have shown to outperform the ARIMA model.

For the computation time, it takes more than 5-hour CPU time on a SunSPARC 5 workstation to develop the ARIMA model and to generate synthesized video traffic of length 171,000[7]. It only takes 3 minutes on the same machine for our algorithm to complete the same tasks.

---

[6] Due to page limit, we only provide plots for JPEG. GOP has similar results and was reported in [6].

[7] The computational complexity to generate synthesized video traffic of length N is $O(N^2)$ for an ARIMA model.

3 Theory

The buffer loss probability has been analyzed to demonstrate the feasibility of wavelet models for theoretical analysis.

Theorem *Let B_N and $\hat{B}_N$ be the buffer sizes at the Nth time slot due to the synthesized traffic by the our wavelet model, and by the FGN process, respectively. Let C and B represent the capacity, and the maximum allowable buffer size respectively. Then*

$$\begin{aligned} \ln Pr(B_N > B) &\sim \ln Pr(\hat{B}_N > B) \\ &\sim \frac{(C-\mu)^2(\frac{B}{C-\mu})^{2(1-H)}(\frac{1-H}{H})^{2H}}{2\sigma^2(1-H)^2}, \end{aligned} \tag{3}$$

where $\frac{1}{2} < H < 1$ is the Hurst parameter. μ and σ^2 is the mean and the variance of the traffic, respectively.

This result[8] demonstrates that using our simple wavelet model which neglects the correlations among wavelet coefficients, a buffer loss rate obtained is similar to that of the original FGN process as given in[9].

4 Conclusions

In this work, we have developed a simple method to obtain wavelet models for video traffic. We have shown that wavelet modeling provides a unified approach to capture both the long-range and the short-range dependence in the video traffic. Since wavelet transforms reduce tremendously the temporal dependence in the traffic, wavelet coefficients can be well modeled as either independent or Markov-dependent random variables. This makes wavelet modeling computationally efficient, and feasible for performance analysis as well. Currently, we are investigating how to using the wavelet models for performance analysis and network control.

References

[1] I. Daubechies, *Ten Lectures on Wavelets.* Philadelphia: SIAM, 1992.

[2] Mark W. Garrett and Walter Willinger. "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic.", in Proceedings of ACM SIGCOMM'94, London, U.K, Aug., 1994

[3] C. Huang, M. Devetsikiotis, I. Lambadaris and A.R. Kaye, "Modeling and Simulation of Self-Similar Variable Bit Rate Compressed Video: A Unified Approach", in Proceedings of ACM SIGCOMM'96, pp. 114-125.

[4] Lance M. Kaplan and C.-C. Jay Kuo. Fractal estimation from noisy data via discrete fractional gaussian noise (dfgn) and the haar basis. *IEEE Transactions on Signal Processing*, Vol. 41, 1993.

[5] Predrag R. Jelenlnovic, Aurel A. Lazar, and Nemo Semret. The effect of multiple time scales and subexponentiality in mpeg video streams on queuing behavior. *IEEE Journal on Selected Area of Communications*, 15, to appear in May 1997.

[6] S. Ma and C. Ji, "Wavelet Modeling of Video Traffic", Tech. Report, Rensselaer Polytechnic Institute, 1997.

[7] B. Melamed, D. Raychaudhuri, B. Sengupta, and J. Zdepski. Tes-based video source modeling for performance evaluation of integrated networks. *IEEE Transactions on Communications*, 10, 1994.

[8] Patrick Flandrin, "Wavelet Analysis and Synthesis of Fractional Brownian Motion", *IEEE transactions on Information Theory*, vol. 38, No.2, pp.910-917, 1992.

---

[8] The proofs are given in [6].

[9] Ilkka Norros, "A storage model with self-similar input," *Queuing Systems*, vol.16, 387-396, 1994.

[10] O. Rose. "Statistical properties of mpeg video traffic and their impact on traffic modeling in atm traffic engineering", Technical Report 101, University of Wurzburg, 1995.

[11] G.W. Wornell and A.V. Oppenheim. wavelet-based representations for a class of self-similar signals with application to fractal modulation, 1992.

Optimal Generator Matrices of Multi-Bit Information Blocks in the Golay Code

Houshou Chen and **John T. Coffey**
Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109, USA

Abstract

We consider the problem of finding an optimal generator matrix of a linear block code $C[n, k]$ over a binary symmetric channel with crossover probability p when the goal of the decoding is to minimize the probability of multi-bit information blocks being in error. We observe that if any two subcodes D_1, D_2 of dimension $k - b$ which are equivalent, i.e., $\exists \pi \in Aut(C)$ such that $\pi(D_1) = D_2$, then they have the same probability of a b-bit information block being in error. The determination of optimal generator matrices of the Golay code for 1-bit, 2-bit, 10-bit, and 11-bit information blocks is examined and we show the systematic generator matrix is optimal in the quiet region, i.e., as $p \to 0$.

1 Introduction

Dunning [3] observes that the probability of an information bit for any given linear block code depends both on the generator matrix and the decoding method. In addition, he gives a greedy algorithm to construct an optimal encoder/decoder pair such that all message symbol error rates are minimized. Kiely et al. [5] present a closed form of optimal decoding strategy for both quiet region ($p \to 0$) and noisy region ($p \to 1/2$) over the binary symmetric channel. Especially, the decoding rule in the noisy region for 1-bit blocks is shown to be very simple: use a separable generator matrix and then decode by ignoring all parity check bits. It was shown by Tolhuizen [7] that the number of computations needed for the determination of an optimal generator matrix of code C to minimize the probability of 1-bit blocks being in error can be reduced by using the code's automorphism group, $Aut(C)$. Here more general results will be extended to a multi-bit information block on any linear block code in section 2 and applied to the investigation of the Golay code. We assume in what follows that the reader is familiar with some definitions of the Golay code, like a sextet or an orbit [1, 6].

In section 3, the optimal subcodes of the Golay code of a 1-bit, 2-bit, 10-bit, and 11-bit information block will be considered in the quiet region, i.e., $p \to 0$. Any two subcodes D_1 and D_2 of C with the same dimension are equivalent, $D_1 \sim D_2$, if there exists an automorphism $\pi \in Aut(C)$, such that $\pi(D_1) = D_2$ and the equivalence class of D is $[D] = \{D' \subset C : \exists \pi \in Aut(C), \pi(D') = D\}$. We find there are only 4, 12, 12, and 4 equivalence classes in the subcodes with dimension 1,2,10, and 11, respectively.

The sextet ([6], p.642) is a set of 6 tetrads with the property that the union of any two tetrads is a codeword of weight 8 in the Golay code; we will derive the intersection of the sextet with any w-set in the Golay code. With this information, the probability of a b-bit information block being in error, $P_{inf,b}(D)$, can be obtained to first order, i.e., the coefficient of the lowest power of p can be determined, and optimal generator matrices of 1-bit, 2-bit, 10-bit, and 11-bit blocks in the quiet region will follow.

2 Multi-Bit Information Block

Here $F_q = GF(q)$ is an arbitrary finite field of order q and F_q^n is a n-dimensional vector space over the finite field F_q. The support of a vector v is the set of nonzero positions in v and the support of a set of vectors is the union of the individual supports. The sum of two codes C_1 and C_2 is defined as

$$C = C_1 + C_2 = \{u + v : u \in C_1, v \in C_2\}.$$

Moreover, if $C_1 \cap C_2 = \{0\}$, then the resulting sum C is called the direct sum of C_1 and C_2, denoted as $C = C_1 \oplus C_2$. Since C is a subspace of F_2^n, the set F_2^n of all vectors can be partitioned into cosets of C:

$$F_2^n = C \cup (q_1 + C) \cup \cdots \cup (q_{2^{n-k}-1} + C),$$

and the set of all cosets of C in F_2^n is denoted F_2^n/C. Similarly, if D is a subcode of C with dimension i, the set C can also be partitioned into cosets of D:

$$C = D \cup (x_1 + D) \cup \cdots \cup (x_{2^{k-i}-1} + D),$$

and C/D is the set of all cosets of D in C.

2.1 Subcodes of C vs. Generator Matrices of C

Let $\Gamma_i = \{D \subset C : \dim(D) = i\}$ be the collection of all subcodes of C with dimension i. The size of Γ_i is known as the binary Gaussian binomial coefficient ([6], p.443), i.e.,

$$|\Gamma_i| = \left[\begin{matrix} k \\ i \end{matrix} \right] = \frac{(2^k - 1)(2^{k-1} - 1) \ldots (2^{k-i+1} - 1)}{(2^i - 1)(2^{i-1} - 1) \ldots (2 - 1)}.$$

For each subcode D of C with dimension $k - i$ there exist i vectors in F_2^n, $E = \{e_1, e_2, \ldots, e_i\}$ which will be called a linear functional representation of D, such that $D^\perp = C^\perp \oplus < E >$, where $< S >$ denotes the code it generates. That is, for each $D \in \Gamma_{k-i}$, a parity check matrix H_D of D can be constructed by adding i more linearly independent parity checks $\{e_1, e_2, \ldots, e_i\}$ into a parity check matrix $H^T = [h_1^T, h_2^T, \ldots, h_{n-k}^T]$ of C:

$$H_D^T = [H^T, e_1^T, \ldots, e_i^T].$$

Therefore, an equivalent representation of $D \in \Gamma_{k-i}$ will be its corresponding linear functional representation $E = \{e_1, e_2, \ldots, e_i\}$.

Any two subcodes D_1, D_2, with $\dim(D_1) = k - k_1$, $\dim(D_2) = k - k_2$ and corresponding linear functionals $E^T = [e_1^T, \ldots, e_{k_1}^T]$, $F^T = [f_1^T, \ldots, f_{k_2}^T]$, is said to be linearly independent if $H \cup E \cup F$ is linearly independent, or equivalently $D_1^\perp + D_2^\perp = C^\perp \oplus < E > \oplus < F >$, i.e., a parity check matrix of $D_1 \cap D_2$ can be chosen as

$$H_{D_1 \cap D_2}^T = [H^T, E^T, F^T].$$

The definition follows if we have a finite collection of subcodes of C.

Definition 1 If $\{D_1, D_2, \ldots, D_m\}$ is a set of m subcodes with $dim(D_i) = k - k_i$, i.e., $D_i \in \Gamma_{k-k_i}$, and corresponding linear functional representations $\{E_1, E_2, \ldots, E_m\}$, we say $\{D_1, D_2, \ldots, D_m\}$ is linearly independent if $H \cup E_1 \cup \ldots \cup E_m$ is linearly independent, or equivalently

$$D_1^\perp + D_2^\perp + \ldots + D_m^\perp = C^\perp \oplus < E_1 > \oplus < E_2 > \oplus \cdots \oplus < E_m > .$$

A generator matrix G of C can be obtained by adding k more linearly independent parity checks $\{f_1, f_2, \ldots, f_k\}$ into a parity check matrix H and taking the inverse [3], i.e.,

$$[f_1, f_2, \ldots f_k, H^T]^{-1} = \begin{bmatrix} G \\ x_1 \\ \vdots \\ x_{n-k} \end{bmatrix}.$$

Therefore, $\{D_1, D_2, \ldots D_m\}$ with $\dim(D_i) = k - k_i$ and corresponding linear functionals $\{E_1, E_2, \ldots E_m\}$ can be used to construct a generator matrix G of C if $\{D_1, D_2, \ldots, D_m\}$ is linearly independent and $\sum_{i=1}^m k_i = k$ or, equivalently,

$$D_1^\perp + D_2^\perp + \cdots + D_m^\perp = C^\perp \oplus < E_1 > \oplus < E_2 > \oplus \cdots \oplus < E_m > = F_2^n.$$

Consequently, if we arrange $E_1 \cup E_2, \cup \ldots, \cup E_m$ such that

$$[E_1^T, E_2^T, \ldots, E_m^T, H^T]^{-1} = \begin{bmatrix} G \\ x_1 \\ \vdots \\ x_{n-k} \end{bmatrix},$$

then D_i is spanned by all row vectors of G by deleting row vectors of G from $\sum_{j=1}^{i-1} k_j + 1$ to $\sum_{j=1}^{i} k_j$. We observe that $[E_1^T, E_2^T, \ldots E_m^T]$ can represent a right inverse G^{-1} of G.

2.2 $P_{inf,b}(D)$: Probability of a b-bit Information Block In Error

For any $e \in F_2^n$ and any $S \subset F_2^n$, we define the probability of e: $P(e) = p^{|e|}(1-p)^{n-|e|}$ and the probability of S: $P(S) = \sum_{c \in S} P(c)$, where p is the crossover probability of a binary symmetric channel. If we use an $[n, k]$ linear block code C with generator matrix G to transmit the information vector u, i.e., the transmitted codeword $c = uG$, then the probability of incorrectly decoding the ith information bit [2, 3, 5] under the optimal decoding rule is

$$P_{inf}^{(i)}(G) = P_{inf,1}(D) = 1 - \sum_{q \in Q} \max[P(q + D), P(q + C - D)],$$

where D is a subcode of C with dimension $k - 1$ and generated by all row vectors of G except the ith one and Q is any set of coset representatives. In other words, with the subcode D fixed, for each coset $q+C$, we are dividing $q+C$ into two cosets of D over $q+C$, i.e., $q+D$ and $q+C-D$, choosing the one with higher probability, and summing over all coset representatives Q. Consequently, a Gale optimal generator matrix G^* [3] among all other generators G, satisfying the constraint

$$P_{inf}^{(1)}(G) \leq P_{inf}^{(2)}(G) \leq \ldots \leq P_{inf}^{(k)}(G),$$

can be constructed by choosing the first k linearly independent subcodes of dimension $k-1$: $\{D_1^*, D_2^*, \ldots D_k^*\}$ with increasingly smallest $P_{inf,1}(D_i^*)$.

Similar results can be developed for multi-bit blocks when the goal is to minimize the probability of a b-bit information block being in error [2]. The probability of a b-bit information block, say the first b bits, being in error under the optimal decoding is

$$\begin{aligned} P_{inf}^{[1,b]}(G) &= P_{inf,b}(D) \\ &= 1-\sum_{q\in Q} \max[P(q+D), P(q+x_1+D), \ldots, P(q+x_{2^b-1}+D)], \end{aligned}$$

where $D \in \Gamma_{k-b}$ and is generated by last $k-b$ row vectors of G. Therefore, for each coset $q+C$, it is divided into 2^b cosets of D in $q+C$ with the one of higher probability is chosen for all entire coset representative Q.

In addition if b is a divisor of k, say $k=b\cdot m$, then in this case the information k bits are divided into m blocks, each with b bits representing some symbol, and we are interested in finding an optimal generator matrix which minimizes simultaneously the probability of each symbol being in error. For example, eight bytes can be encoded using a [128,64] BCH code, and we would like to know how to find an optimal generator matrix in order to minimize the probability that bytes are in error. By choosing the first $m = k/b$ linearly independent subcodes of dimension $k-b$: $\{D_1^*, D_2^*, \ldots, D_m^*\}$ with increasingly smallest $P_{inf,b}(D_i^*)$, we can obtain a Gale optimal generator matrix G^* among all other generators G, satisfying the constraint

$$P_{inf}^{[1,b]}(G) \leq P_{inf}^{[b+1,2b]}(G) \leq \ldots \leq P_{inf}^{[(m-1)b+1,mb]}(G).$$

Moreover, if $\{E_1^*, E_2^*, \ldots, E_m^*\}$ are linear functionals of $\{D_1^*, D_2^*, \ldots, D_m^*\}$, then any E_i' obtained by row operations on E_i is still a candidate linear functional of D_i. This implies if G^* is a Gale optimal generator matrix of b-bit information blocks of C, then any G' obtained by row operations of G^* within each block is still a Gale optimal generator matrix. In conclusion, the optimal subcode D^* with the smallest $P_{inf,b}(D^*)$ will be our concern.

2.3 Equivalence Relation on Γ_i

We will show if $D_1, D_2 \in \Gamma_{k-b}$ and $D_1 \sim D_2$, then we have $P_{inf,b}(D_1) = P_{inf,b}(D_2)$. Therefore, to find an optimal subcode $D^* \in \Gamma_{k-b}$, the number of comparisons can be considerably reduced to the number of equivalence classes instead of order of Γ_{k-b}.

Lemma 1 *Let π be any permutation in $Aut(C)$ such that $\pi(D_1) = D_2$. Two codewords of C, x_1 and x_2, are in the same coset of C/D_1 if and only if $\pi(x_1)$ and $\pi(x_2)$ are in the same coset of C/D_2.*

Proof. If x_1 and x_2 are in the same coset of D_1 in C, then we have

$$\begin{aligned} & x_1 + D_1 = x_2 + D_1 \\ \iff & x_1 - x_2 \in D_1 \\ \iff & \pi(x_1 - x_2) \in \pi(D_1) = D_2 \\ \iff & \pi(x_1) - \pi(x_2) \in D_2 \\ \iff & \pi(x_1) + D_2 = \pi(x_2) + D_2. \end{aligned} \qquad \square$$

With similar reasoning, we have the following Lemma also.

Lemma 2 *Let π be any permutation in $Aut(C)$. Any two words of F_2^n, q_1 and q_2, are in the same coset of F_2^n/C if and only if $\pi(q_1)$ and $\pi(q_2)$ are in the same coset of F_2^n/C.*

In the binary symmetric channel, $P(e) = p^{|e|}(1-p)^{n-|e|}$ depends only on the weight of e and not on 1's positions in e. In other words, we always have that $P(\pi(e)) = P(e)$ for any permutation $\pi \in S_n$ and any vector $e \in F_2^n$. Applying π with $\pi(D_1) = D_2$ to the 2^b cosets of D_1 in C, the weight distributions on the 2^b cosets of D_1 in C over $q + C$:

$$\{q + D_1, q + x_1 + D_1, \ldots, q + x_{2^b-1} + D_1\}$$

is the same as the weight distributions on the 2^b coset of D_2 in C over $\pi(q) + C$:

$$\{\pi(q) + D_2, \pi(q) + \pi(x_1) + D_2, \ldots, \pi(q) + \pi(x_{2^b-1}) + D_2\}.$$

Moreover, by Lemma 1 if $X = \{x_0, x_1, x_2, \ldots, x_{2^b-1}\}$ is a coset representative of C/D_1, then $\pi(X) = \{\pi(x_0), \pi(x_1), \pi(x_2), \ldots, \pi(x_{2^b-1})\}$ is also a coset representative of C/D_2 and we have

$$\begin{aligned} \max_{x \in X}\{P(q + x + D_1)\} &= \max_{i=0}^{2^b-1}\{P(q + x_i + D_1)\} \\ &= \max_{i=0}^{2^b-1}\{P(\pi(q) + \pi(x_i) + D_2)\} \\ &= \max_{x \in \pi(X)}\{P(\pi(q) + x + D_2)\}. \end{aligned}$$

Similarly, by Lemma 2, if Q is a coset representative of F_2^n/C, then $\pi(Q)$ is another coset representative. Thus, we have

$$\begin{aligned} P_{inf,b}(D_1) &= 1 - \sum_{q \in Q} \max_{x \in X}\{P(q + x + D_1)\} \\ &= 1 - \sum_{q \in Q} \max_{x \in X}\{P(\pi(q) + \pi(x) + \pi(D_1))\} \\ &= 1 - \sum_{q \in Q} \max_{x \in \pi(X)}\{P(\pi(q) + x + D_2))\} \\ &= 1 - \sum_{q \in \pi(Q)} \max_{x \in \pi(X)}\{P(q + x + D_2))\} \\ &= P_{inf,b}(D_2). \end{aligned}$$

3 Optimal Subcodes of Multi-Bit Information Block in the Golay Code

In section 3, C=$\mathcal{C}_{24}$ is the binary Golay code of length 24, which is obtained by adding an overall parity check to the cyclic Golay code $\mathcal{C}_{23}$ whose generator polynomial is 1 +

$\mathbf{a}(x, L_4)$	$b(x \in S_1)$	$\mathbf{a}(x, L_4)$	$b(x \in S_2)$	$\mathbf{a}(x, L_4)$	$b(x \in S_3)$
(5,1,0,0,0)	1771	(5,0,1,0,0)	231	(5,0,0,1,0)	21
		(4,2,0,0,0)	1540	(4,1,1,0,0)	630
				(3,3,0,0,0)	1120

Table 1: $x \in S_1$, S_2, or S_3.

$x^2 + x^4 + x^5 + x^6 + x^{10} + x^{11}$. The automorphism group of C_{24} is the Mathieu group M_{24}, which is five-fold transitive [1, 6]. Let the set $Q_i = \{x \in F_2^{24} : |x| \leq 3\}$, $1 \leq i \leq 3$ and $Q_4 = \{x \in F_2^{24} : |x| = 4 \text{ and } x_1 = 1\}$, then $Q_1 \cup Q_2 \cup Q_3 \cup Q_4$ is a complete set of coset representatives of $C = C^{\perp}$. As described in [1], there are 49 orbits of vectors in F_2^{24} under the action of M_{24}, denoted by S_w, T_w, U_w, P_{12} and X_{12}, where the subscript gives the weight of the vectors. In the following, O_w is used to represent the orbits of the vectors in F_2^{24} or the set of the support of any vector in O_w.

We will restrict ourselves to the quiet region, i.e., sufficiently small crossover probability p. When the coset $q + C$ of C has a unique minimum weight element, the coset $q + x + D$ of D with higher probability is the one containing this minimum weight. Since the Golay code has unique coset leader of weight up to 3, therefore the lowest power of p with nonzero coefficient in $P_{inf,b}(D)$ of D is p^4. In other word, to find an optimal subcode D^* of a b-bit information block, we need to compare the coefficient of p^4 of $P_{inf,b}(D)$ for all subcodes $D \in \Gamma_{12-b}$ of C. Any coset $q + C$ of C with a coset leader of weight 4 contains a sextet and there are $\binom{23}{3} = 1771$ sextets in the Golay code. We will derive the intersection of these sextets with any w-set and the coefficient of p^4 in $P_{inf,b}(D)$ will be determined with this information.

3.1 Intersection of w-set with the Six Tetrads of a Sextet

The set of all 1771 sextets will be denoted as V. For each sextet $L_4 \in V$ and each vector $x \in F_2^{24}$ we define the vector $\mathbf{a}(x, L_4) = (a_0, a_1, a_2, a_3, a_4)$ [7] where a_i denotes the number of tetrads in L_4 that has i ones on the nonzero positions of x. By the definition of orbit [1], if x_1 and x_2 are in the same orbit, i.e., there exists a $\pi \in Aut(C)$ such that $\pi(x_1) = x_2$, then we have $\mathbf{a}(x_1, L_4) = \mathbf{a}(\pi(x_1), \pi(L_4)) = \mathbf{a}(x_2, \pi(L_4))$ and when L_4 runs entirely over V, $\pi(L_4)$ also runs entirely over V. Therefore, the different types of $\mathbf{a}(x_1, L_4)$ over all $L_4 \in V$ are the same as the different types of $\mathbf{a}(x_2, L_4)$ over all $L_4 \in V$ and it is sufficient to consider those vectors x's in different orbits. By $b_j(x)$ we denote the number of sextets that are of type $\mathbf{a}_j$ with respect to x, i.e.,

$$b_j(x) = |\{L_4 \in V : \mathbf{a}(x, L_4) = \mathbf{a}_j\}.$$

The possible different types of $\mathbf{a}(x, L_4)$ over all $L_4 \in V$ for x in different orbits up to 6-set are in the table 1-2.

3.2 $P_{inf,1}(D)$ and $P_{inf,2}(D)$

b=1: There are only four equivalence classes on Γ_{11} since M_{24} is 5-fold transitive: D_1 with $E_1 = \{(10^{23})\}$, D_2 with $E_2 = \{(1^2 0^{22})\}$, D_3 with $E_3 = \{(1^3 0^{21})\}$, and D_4 with $E_4 =$

$\mathbf{a}(x, L_4)$	$b(x \in S_4)$	$\mathbf{a}(x, L_4)$	$b(x \in S_5)$	$\mathbf{a}(x, L_4)$	$b(x \in S_6)$	$b(x \in U_6)$
(5,0,0,0,1)	1	(4,1,0,0,1)	5	(4,0,1,0,1)	15	0
(4,1,0,1,0)	80	(4,0,1,1,0)	30	(3,2,0,0,1)	0	15
(4,0,2,0,0)	960	(3,2,0,1,0)	160	(4,0,0,2,0)	20	0
(3,2,1,0,0)	90	(3,1,2,0,0)	360	(3,1,1,1,0)	0	180
(2,4,0,0,0)	640	(2,3,1,0,0)	960	(2,3,0,1,0)	320	180
		(1,5,0,0,0)	256	(3,0,3,0,0)	120	45
				(2,2,2,0,0)	720	675
				(0,6,0,0,0)	96	46
				(1,4,1,0,0)	480	630

Table 2: $x \in S_4$, S_5, S_6, or U_6.

$\{(1^4 0^{20})\}$. To find the coefficient of p^4 in $1 - P_{inf,1}(D)$ with a linear functional $E \in O_w$, compare the number of six tetrads contained in between two sets $\{x \in L_4 : x \cdot E = 0\}$ and $\{x \in L_4 : x \cdot E = 1\}$ over V, choose the one with higher value, and sum them up. The corresponding probability of ith information bit being in error $P_{inf,1}(D)$ is :

$$\begin{aligned}
P_{inf,1}(D_1) &= (1 \cdot b_1(x \in S_1))p^4 + o(p^5) = 1771p^4 + o(p^5), \\
P_{inf,1}(D_2) &= (0 \cdot b_1(x \in S_2) + 2 \cdot b_2(x \in S_2))p^4 + o(p^5) = 3080p^4 + o(p^5), \\
P_{inf,1}(D_3) &= (1 \cdot b_1(x \in S_3) + 1 \cdot b_2(x \in S_3) + 3 \cdot b_3(x \in S_3))p^4 + o(p^5) \\
&= 4011p^4 + o(p^5), \\
P_{inf,1}(D_4) &= (0 \cdot b_1(x \in S_4) + 2 \cdot b_2(x \in S_4) + 2 \cdot b_3(x \in S_4) \\
&\quad +0 \cdot b_4(S_4) + 2 \cdot b_5(S_4))p^4 + o(p^5) = 3360p^4 + o(p^5).
\end{aligned}$$

An optimal systematic generator matrix is then constructed by choosing 12 linearly independent subcodes D_is with $P_{inf}(D_i) = 1771p^4 + o(p^5)$.

b=2: The possible linear functionals of $D \in \Gamma_{10}$ is constructed by combining any pair of vectors in $Q_1 \cup Q_2 \cup Q_3 \cup Q_4$ and by Fig. 1 in [1] and properties of the Golay code, we find there are 12 equivalence classes on Γ_{10}. Compare the number of six tetrads contained in between four sets: $\{x \in L_4 : x \cdot E = (0,0)\}$, $\{x \in L_4 : x \cdot E = (0,1)\}$, $\{x \in L_4 : x \cdot E = (1,0)\}$, and $\{x \in L_4 : x \cdot E = (1,1)\}$, the coefficient of p^4 in $1 - P_{inf,2}(D)$ can be obtained like $1 - P_{inf,1}(D)$ by Table 1-2. For example, $1 - P_{inf,2}(D) = (5 \cdot b_1(S_2) + 4 \cdot b_2(S_2))p^4 + o(p^5)$ $= 7315p^4 + o(p^5)$ if the corresponding linear functional E of D is $\{(10^{23}), (010^{22})\}$. The probability of a 2-bit information block being in error $P_{inf,2}(D)$ is:

$$\begin{aligned}
P_{inf,2}(D_1) &= 3311p^4 + o(p^5) \text{ with } E_1 = \{(10^{23}), (010^{22})\}, \\
P_{inf,2}(D_2) &= 4431p^4 + o(p^5) \text{ with } E_2 = \{(10^{23}), (01^2 0^{21})\}, \\
P_{inf,2}(D_3) &= 4571p^4 + o(p^5) \text{ with } E_3 = \{(10^{23}), (01^3 0^{20})\}, \\
P_{inf,2}(D_4) &= 4620p^4 + o(p^5) \text{ with } E_4 = \{(1^2 0^{22}), (01^2 0^{21})\}, \\
P_{inf,2}(D_5) &= 5400p^4 + o(p^5) \text{ with } E_5 = \{(1^2 0^{22}), (0^2 1^2 0^{20})\}, \\
P_{inf,2}(D_6) &= 5551p^4 + o(p^5) \text{ with } E_6 = \{(1^2 0^{22}), (01^3 0^{20})\}, \\
P_{inf,2}(D_7) &= 5511p^4 + o(p^5) \text{ with } E_7 = \{(1^2 0^{22}), (0^2 1^3 0^{19})\}, \\
P_{inf,2}(D_8) &= 5476p^4 + o(p^5) \text{ with } E_8 = \{(1^2 0^{22}), (01^4 0^{19})\}, \\
P_{inf,2}(D_9) &= 5691p^4 + o(p^5) \text{ with } E_9 = \{(1^3 0^{21}), (0^2 1^3 0^{19})\}, \\
P_{inf,2}(D_{10}) &= 5691p^4 + o(p^5) \text{ with } E_{10} = \{(1^3 0^{21}), (0^3 1^3 0^{18})\},
\end{aligned}$$

$$\begin{aligned}
P_{inf,2}(D_{11}) &= 5680p^4 + o(p^5) \quad \text{with} \quad E_{11} = \{(1^4 0^{20}), (0^2 1^4 0^{18})\}, \\
P_{inf,2}(D_{12}) &= 5808p^4 + o(p^5) \quad \text{with} \quad E_{12} = \{(1^4 0^{20}), (0^2 1^2 0 1^2 0^{17})\}.
\end{aligned}$$

Note E_i is the corresponding linear functional of D_i and the support of E_{10} and E_{11} is in U_6, while the support of E_{12} is in S_6. Hence, choosing six linearly independent subcodes D_i's with $P_{inf,2}(D_i) = 3311p^4 + o(p^5)$ assures a systematic generator matrix is optimal if pair-by-pair decoding is concerned.

3.3 $P_{inf,10}(D)$ and $P_{inf,11}(D)$

Lemma 3 *If a subcode $D \in \Gamma_{12-b}$ of the Golay code C contains no codewords of weight 8, then $P_{inf,b}(D) = 8855p^4 + o(p^5)$ and if D only contains one codeword of weight 8, then we have $P_{inf,b}(D) = 8820p^4 + o(p^5)$.*

Proof. If a subcode $D \in \Gamma_{12-b}$ contains no codewords of weight 8, then any coset $q + D$ in F_2^n/D cannot contain more than one tetrad since the union of any two tetrads in the same coset $q + C$ in F_2^n/C form an octad in the subcode D. Thus, we have $1 - P_{inf,b}(D) = 1771p^4 + o(p^5)$ and $P_{inf,b}(D) = (\binom{24}{4} - 1771)p^4 + o(p^5) = 8855p^4 + o(p^5)$. If a subcode $D \in \Gamma_{12-b}$ contains only one octad, then any coset $q + D$ in F_2^n/D contains no more than two tetrads. Among $\{q + D, q + x_1 + D, \ldots, q + x_{2^b-1} + D\}$, there can be only one of them containing two tetrads, say $q + x_i + D$, and we can always choose another q' such that $q' + D$ is the only one containing two tetrads among $\{q' + D, q' + x_1 + D, \ldots, q' + x_{2^b-1} + D\}$. Therefor, there are only $\frac{1}{2}\binom{8}{4}$ cosets such that $q + D$ has two tetrads among $q \in Q_4$ and we have $1 - P_{inf,b}(D) = (1771 + \frac{1}{2}\binom{8}{4})p^4 + o(p^5)$, i.e., $P_{inf,b}(D) = (8855 - \frac{1}{2}\binom{8}{4})p^4 + o(p^5) = 8820p^4 + o(p^5)$. □

b=11: The possible weight enumerators of subcode $D \in \Gamma_1$ are $1+x^8$, $1+x^{12}$, $1+x^{16}$, and $1+x^{24}$. Since codewords of the same weight of the Golay code are in the same orbit, this implies there are only four equivalence classes on Γ_1 and by lemma 3 the probability of a 11-bit information block being in error $P_{inf,11}(D)$ is:

$$\begin{aligned}
P_{inf,11}(D) &= 8820p^4 + o(p^5) \quad \text{if} \quad W_D(x) = 1 + x^8 \\
P_{inf,11}(D) &= 8855p^4 + o(p^5) \quad \text{if} \quad W_D(x) = 1 + x^{12} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\; = 1 + x^{16} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\; = 1 + x^{24}.
\end{aligned}$$

The support of the corresponding linear functional of subcode D with $W_D(x) = 1 + x^8$ is in T_{11}, therefore a systematic generator matrix with the 11 desired bits in T_{11} is optimal to minimize the error probability of this 11-bit block.

b=10: The possible weight enumerators $W_D(x)$ of $D \in \Gamma_2$ can be obtained by choosing any pair of nonzero codewords of C as row vectors of generator matrix of D and it turns out there are only 12 possible weight enumerators. With the properties of the Golay code, those subcodes with the same weight enumerator are all equivalent and the probability of a 10-bit information block being in error $P_{inf,10}(D)$ is :

$$\begin{aligned}
P_{inf,10}(D) &= 8751p^4 + o(p^5) \quad \text{if} \quad W_D(x) = 1 + 3x^8 \\
P_{inf,10}(D) &= 8785p^4 + o(p^5) \quad \text{if} \quad W_D(x) = 1 + 2x^8 + x^{12} \\
P_{inf,10}(D) &= 8792p^4 + o(p^5) \quad \text{if} \quad W_D(x) = 1 + 2x^8 + x^{16}
\end{aligned}$$

$$\begin{aligned}
P_{inf,10}(D) &= 8820p^4 + o(p^5) \quad \text{if} \quad W_D(x) &&= 1 + x^8 + 2x^{12} \\
& && = 1 + x^8 + x^{12} + x^{16} \\
& && = 1 + x^8 + 2x^{16} \\
& && = 1 + x^8 + x^{16} + x^{24} \\
P_{inf,10}(D) &= 8855p^4 + o(p^5) \quad \text{if} \quad W_D(x) &&= 1 + 3x^{12} \\
& && = 1 + 2x^{12} + x^{16} \\
& && = 1 + 2x^{12} + x^{24} \\
& && = 1 + x^{12} + 2x^{16} \\
& && = 1 + 3x^{16}.
\end{aligned}$$

The support of the corresponding linear functional of subcode D with $W_D(x) = 1 + 3x^8$ is in T_{10}, therefore a systematic generator matrix with the 10 desired bits in T_{10} is optimal if a 10-bit information block is our concern.

4 Conclusion

We have considered the problem of determining the optimal subcodes for the Golay code to minimize the probability of a 1-bit, 2-bit, 10-bit, or 11-bit block being in error. The problem remains to be solved is the case of blocks of 3 to 9 bits. We also have results on the binary Hamming codes: to minimize simultaneously the error probability of an information bit in the quiet region, any separable generator matrix is optimal.

References

[1] J. H. Conway and N. J. A. Sloane, "Orbit and coset analysis of the Golay and related codes," IEEE Trans. Inform. Theory, vol. 36, no. 5, pp. 1038-1050, Sept. 1990.

[2] P. Delsarte, "Partial-optimal piecewise decoding of linear codes," IEEE Trans. Inform. Theory, vol. IT-24, no. 1, pp. 70-75, Jan. 1978.

[3] L. A. Dunning, "Encoding and decoding for the minimization of message symbol error rates in linear block codes," IEEE Trans. Inform. Theory, vol. IT-33, no. 1, pp. 91-104, Jan. 1987.

[4] M. J. Erickson, Introduction to Combinatorics, Wiley, New York.

[5] A. B. Kiely, J. T. Coffey and M. R. Bell, "Optimal information bit decoding of linear block codes," IEEE Trans. Inform. Theory, vol. 41, no. 1, pp. 130-140, Jan. 1995.

[6] F. J. MacWilliams and N. J. Sloane, The Theory of Error-Correcting Codes, New York: North-Holland, 1978.

[7] L. M. G. M. Tolhuizen and W. J. van Gils, "A large automorphism group decreases the number of computations in the construction of an optimal encoder/decoder pair for a linear block code," IEEE Trans. Inform. Theory, vol, 34, no. 2, pp. 333-338, Mar. 1988.

Binomial Bounds on Spectra of Codes with Known Dual Distance

Ilia Krasikov
Tel-Aviv University
School of Mathematical Sciences,
Ramat-Aviv 69978 Tel-Aviv, Israel and
Beit-Berl College, Kfar-Sava, Israel

Simon Litsyn*
Tel-Aviv University
Department of Electrical Engineering – Systems
Ramat-Aviv 69978
Tel-Aviv, Israel

Abstract

For codes with known dual distance we estimate the interval where the distance distribution is simultaneously upper- and lower-bounded by the binomial distribution. The lower bound means that for every subinterval of size about $\sqrt{n \ln n}$ there exists a component asymptotically achieving the binomial bound.

1. Introduction

It is well-known (see, e.g. [5, p.287]) that typical random linear binary codes have the binomial distance distribution (spectrum). Namely, the expected number of codewords of weight $i > 0$ in such a code C of length n is clearly $\binom{n}{i}|C|/2^n$. It is also known that accuracy of the binomial approximation increases with the dual distance of code. Let d' be the dual distance of a code, and $\delta' = d'/n$. In [2] we have shown that the distance distribution is asymptotically in n upperbounded by the binomial distribution in the range

$$\left(\frac{n}{2}\left(1-\sqrt{\delta'(2-\delta')}\right), \frac{n}{2}\left(1+\sqrt{\delta'(2-\delta')}\right)\right).$$

For even codes the binomial upper bound is valid in the interval

$$\left(\frac{n}{2}\left(1-2\delta'\right)^2, n-\frac{n}{2}\left(1-2\delta'\right)^2\right).$$

In this paper we demonstrate that actually every even code (i.e. a code with only even distances) in the interval

$$\left(\frac{n}{2}\left(1-\frac{\delta'}{1-\delta'}\right), \frac{n}{2}\left(1+\frac{\delta'}{1-\delta'}\right)\right)$$

has an asymptotically binomial distance distribution in the following sense: all the components in this range are upperbounded by const $\cdot \sqrt{n}\binom{n}{i}|C|/2^n$, and for every subinterval of size about $\sqrt{n \ln n}$ there exists a component asymptotically achieving the upper bound.

2. Basic identities

Let $C = C(n, |C|, d)$ be a code of length n, size $|C|$ and with minimum distance d. Let $\mathbf{B} = (B_0, B_1, \dots, B_n)$ stand for the distance distribution of the code C, and $\mathbf{B}' = (B'_0, B'_1, \dots, B'_n)$ stand for the dual distance distribution. We say that a distance distribution is even if its components are nonzero only for even indices. $\mathbf{B}$ and $\mathbf{B}'$ are related by the MacWilliams transform

$$|C'|B_i = \sum_{j=0}^{n} B'_j P_i(j), \tag{1}$$

where $|C'| = \sum_{i=0}^{n} B'_i = 2^n/|C|$, and $P_i(j)$ are Krawtchouk polynomials defined by the following recurrence (see, e.g. [4, 5])

$$(k+1)P_{k+1}(x) = (n-2x)P_k(x) - (n-k+1)P_{k-1}(x), \tag{2}$$

$$P_0(x) = 1, \quad P_1(x) = n - 2x,$$

$$P_{n-i}(x) = (-1)^x P_i(x), \quad \text{for integer } x \in [0, n]. \tag{3}$$

Let D be the maximal index, less than n, such that $B_D > 0$. Let d', the dual distance of C, be the minimal positive index of nonzero component of $\mathbf{B}'$, and D' be defined similarly to D. If all nonzero components in $\mathbf{B}$ are of even indices, such a code is called *even*.

The following lemma is due to Ph.Delsarte [1] (see also [4]).

Lemma 1 *Let* $\alpha(x) = \sum_{i=0}^{r} \alpha_i P_i(x)$, *then*

$$\alpha_0|C'| + |C'|\sum_{i=d}^{r} \alpha_i B_i = \alpha(0) + \alpha(n)B'_n + \sum_{j=d'}^{D'} \alpha(j)B'_j. \tag{4}$$

Proof. Calculating $|C'|\sum_{i=0}^{r} \alpha_i B_i$, we get the claim from (1).

□

Using (3) we analogously have the following lemma.

Lemma 2 *Let* $\alpha(x) = \sum_{i=0}^{r} \alpha_i P_i(x)$, *then, for* n *even,*

$$\alpha_0|C'| + |C'|\sum_{i=d}^{r} \alpha_i B_{n-i} = \alpha(0) + \alpha(n)B'_n + \sum_{j=d'}^{D'} (-1)^j \alpha(j)B'_j. \tag{5}$$

□

Consider now the following polynomials of degree $2k$

$$g_k(x) = \prod_{i=0}^{k-1}((n-2x)^2 - (2i)^2), \tag{6}$$

$$f_k(x) = \prod_{i=0}^{k-1}((n-2x)^2 - (4i)^2), \tag{7}$$

$$h_k(x) = \prod_{i=0}^{k-1}((n-2x)^2 - (4i+2)^2), \tag{8}$$

$g_k(x) = 0$ for all $x = \frac{n}{2} \pm i$, $f_k(x) = 0$ for all $x = \frac{n}{2} \pm 2i$, and $h_k(x) = 0$ for all $x = \frac{n}{2} \pm (2i+1)$, where $i = 0, \ldots, k-1$.

From now on we restrict ourselves to the case n even, $n = 2m$.

Lemma 3

$$g_k(x) = (2k)! \sum_{i=0}^{k} \binom{m-i}{k-i} P_{2i}(x),$$

$$f_k(x) = (2k)! \sum_{i=0}^{k} \frac{m-2i}{m-k-i} \binom{m-k-i}{k-i} P_{2i}(x),$$

$$h_k(x) = (2k)! \sum_{i=0}^{k} \binom{m-k-i-1}{k-i} P_{2i}(x).$$

Proof. The proof is by induction on k using the recurrence relation. We omit details.

□

Theorem 1 *Let C be an even code, then for $k = 1, \ldots, m$,*

$$|C'|\binom{m}{k} + |C'| \sum_{i=d/2}^{k} \binom{m-i}{k-i} B_{2i} =$$

$$2^{2k+1} \frac{m}{m+k} \binom{m+k}{2k} + 2^{2k+1} \sum_{j=d'}^{m-k} \frac{m-j}{m-j+k} \binom{m-j+k}{2k} B'_j.$$

Proof. If C has only even weights, then $B'_n = 1$ and $B'_i = B'_{n-i}$. In Lemma 1 we choose $\alpha(x) = g_k(x)$. Then

$$(2k)!|C'| \sum_{i=0}^{k} \binom{m-i}{k-i} B_{2i} = \sum_{j=0}^{n} B'_j \prod_{i=0}^{k-1}((n-2j)^2 - 4i^2) =$$

$$|C'|\binom{m}{k} + |C'| \sum_{i=d/2}^{k} \binom{m-i}{k-i} B_{2i},$$

and the result follows. □

3. Bounds on spectrum components

To use the identities derived in the previous section we choose k to be such that one of the sums vanishes. It corresponds to the choices of $k < d'/2$ or $k > m - d'$, depending on which sum we like to cancel. The sum left can be lowerbounded just by the maximal term, thus getting an upper bound for the terms in the sum.

We start with the case of arbitrary even codes, $d \geq 4$. Notice, that although our method does not give estimates for $B_{n/2}$, clearly $B_{n/2} < |C|$. Due to the symmetry, about $n/2$, of the dual distance distribution we may restrict ourselves to the components B'_j, $j \leq n/2$. For $k < d/2$ from Theorem 1 we get that for any $j \in [d', D']$, $j \leq m - k$,

$$B'_j < \frac{|C'|\binom{m}{k}(m-j+k)}{2^{2k+1}\binom{m-j+k}{2k}(m-j)}. \tag{9}$$

Choose k to be the nearest integer to $\frac{(m-j)^2}{m}$ (the optimal choice is $\frac{2(m-j)^2-m}{2m-1}$, but it makes the calculations cumbersome, not gaining a lot). To satisfy the restriction $k < d/2$ we require

$$\frac{(m-j)^2}{m} + \frac{1}{2} \leq \frac{d}{2},$$

that is valid for

$$j \geq m - \sqrt{\frac{(d-1)m}{2}}.$$

Theorem 2 *For sufficiently large n and*

$$j \in \left[\frac{n}{2} - \tfrac{1}{2}\sqrt{(d-1)n}, \frac{n}{2} + \tfrac{1}{2}\sqrt{(d-1)n}\right], \; j \neq n/2,$$

$$B'_j < \sqrt{\frac{\pi j(n-j)}{n}}\,\frac{\binom{n}{j}}{|C|}(1 + O(\frac{1}{n}));$$

Otherwise,

$$B'_j < \frac{|C'|\binom{n/2}{d/2-1}(n+d-2j-2)}{2^{d-1}\binom{n/2+d/2-j-1}{d-2}(n-2j)}.$$

Proof. Follows from the previous arguments and Stirling's approximation. The second inequality holds unconditionally for the choice $k = d/2 - 1$. □

To upperestimate B_{2i} in the code we may choose $k \geq m - d' + 1$, $k < m$, and get

$$B_{2i} < \frac{2^{2k+1}m\binom{m+k}{2k}}{|C'|(m+k)\binom{m-i}{k-i}}.$$

Observe that the same estimate holds for $B_{2(m-i)}$. Indeed from Lemma 2 the identity analogous to that of Theorem 1 holds also for B_{2i} with $i > m/2$. The only difference in the two identities is that then we get additional multiplier $(-1)^j$ in the sum of RHS. Since this sum vanishes due to the the choice of k, we arrive at the same inequality.

Optimal choice of k is the nearest integer to $(2mi-2m+1)/(2m-2i-1)$. For simplicity, we choose k to be the nearest integer to $mi/(m-i)$. To satisfy restrictions on k it is necessary that

$$m > \frac{mi}{m-i} - \frac{1}{2} \geq m - d' + 1,$$

yielding

$$\frac{m(2m-2d'+3)}{(4m-2d'+3)} \leq i < m.$$

Substituting the chosen value of k we get using Stirling's approximation

$$B_{2i} < \sqrt{\frac{2i}{m}} \frac{m^{2m}}{|C'|(m-i)^{2(m-i)} i^{2i}}\left(1+O\left(\frac{1}{n}\right)\right).$$

Theorem 3 *For*

$$i \in \left(\frac{n}{4}\left(1-\frac{2d'-3}{2n-2d'+3}\right), \frac{n}{4}\left(1+\frac{2d'-3}{2n-2d'+3}\right)\right), \; i \neq n/2,$$

$$B_{2i} < \frac{4i\sqrt{\pi(n-2i)}\binom{n}{2i}}{n|C'|}\left(1+O\left(\frac{1}{n}\right)\right).$$

Otherwise,

$$B_{2i} < \frac{2^{n-2d'+2} n \binom{n-d'+1}{d'-1}}{|C'|(n-d'+1)\binom{n/2-i}{d'-1}}.$$

□

Now we will show that the binomial upperbound is in fact asymptotically achieved by some spectral components, more rigorously, such a component B_{2i} always exists on any subinterval of size about $\sqrt{i \log n}$. Note, that from the results of Ward [6] it follows that there are no codes of nonzero rate with all the gaps in the distance distribution being of size growing with n.

We formulate the result only for the case of even code. Similar results are valid for the other cases.

Theorem 4 *For sufficiently large n and some appropriate constants c, c_1, c_2 and $d' < n/2 - c_1 \ln n$, there exists $j \in [i - c\sqrt{i \ln n}, i + c\sqrt{i \ln n}]$ such that*

$$B_{2j} \geq \frac{c_2}{n|C'|}\sqrt{\frac{(n-2j)j}{\ln n}}\binom{n}{2j},$$

provided

$$\frac{n}{4}\left(1-\frac{2d'-3}{2n-2d'+3}\right) \leq i \leq \frac{n}{4}\left(1+\frac{2d'-3}{2n-2d'+3}\right).$$

Proof. Fix i, $m(2m-2d'+3)/(4m-2d'+3) \le i < m$, and choose $k = im/(m-i)$. We will make use of the identity of Theorem 1. Using the Plotkin upper bound $|C'| \le d'2^{n-2d'+2}$, it is easy to see that for m growing and $d' < n/2 - c_1 \ln n$,

$$|C'|\binom{m}{k} = o\left(2^{2k+1}\frac{m}{m+k}\binom{m+k}{2k}\right), \tag{10}$$

provided $m - d' + 1 \le k \le m$, and $d' \ge 3$.

Now choose $\epsilon = c\sqrt{\ln m/i}$, where c is an absolute constant.

First we show that for appropriate choice of c and sufficiently large m

$$|C'| \sum_{j,\,|j-i|\ge c\sqrt{i\ln m}} \binom{m-j}{k-j} B_{2j} < 2^{2k}\frac{m}{m+k}\binom{m+k}{2k},$$

that is

$$\sum_{j,\,|j-i|\ge c\sqrt{i\ln m}} \frac{(m+k)|C'|\binom{m-j}{k-j}B_{2j}}{m2^{2k}\binom{m+k}{2k}} = \sum_j S_j = \Sigma < 1.$$

Substitute now the upper estimate for B_{2j} from the previous theorem. The exponentially big part of the summands depending on j is either

$$\binom{m-j}{k-j}\binom{2m}{2j}$$

or

$$\binom{m-j}{k-j}\Big/\binom{m-j}{d'}, \tag{11}$$

depending on whether j belongs to the interval defined in the theorem. In the first case considering the ratio between two consecutive j's, and taking into account that $k = im/(m-i)$, we conclude that the function achieves the only maximum at $j = i$. In the second case, proceeding similarly, we get that (11) is monotonically increasing in j. Hence, the maximal term in the sum occurs at j closest to i. So, denoting $\varepsilon = c\sqrt{\ln m/i}$,

$$\Sigma < k(S_{i(1+\epsilon)} + S_{i(1-\epsilon)}).$$

Using Stirling's approximation and upperestimating B_{2j} as in Theorem 3 we get

$$\ln S_j < 2m\ln m - (k-j)\ln(k-j) - 2j\ln j$$

$$-(m+k)\ln(m+k) - (m-j)\ln(m-j) + 2k\ln k + O(\ln m).$$

So,

$$\ln S_{i(1+\varepsilon)} = -2(1+\varepsilon)i\ln(1+\varepsilon)$$

$$+(i+\varepsilon i - m)\ln\left(1 - \frac{\varepsilon i}{m-i}\right) - \frac{i(i-\varepsilon(m-i))}{m-i}\ln\left(1 - \frac{\varepsilon(m-i)}{i}\right).$$

Applying elementary inequalities for $0 < x < 1$

$$\ln(1-x) > x/(x-1), \quad \ln(1+x) > 2x/(2+x),$$

we obtain

$$\ln S_{i(1+\epsilon)} <$$

$$-\frac{2\epsilon^2 i}{2+\epsilon} + O(\ln m) = -\frac{2c^2 \ln m}{2+\epsilon} + O(\ln m).$$

Similarly,

$$\ln S_{i(1-\epsilon)} <$$

$$-\frac{2i(m^2 - 2im + 2i^2 - \varepsilon i^2 + \varepsilon im)}{(2m - 2i + \varepsilon i)(2i - \varepsilon i + m\varepsilon)} + O(\ln m)$$

$$= -\frac{\varepsilon^2(m^2 - 2im + 2i^2)}{2(m-i)} + O(\ln m).$$

Since, for $i > \ln m$, that is for $d' < m - c_1 \ln m$, $\ln \Sigma$ can be made negative by an appropriate choice of c we conclude that $\Sigma < 1$. Denote $I = [i - c\sqrt{i \ln m}, i + c\sqrt{i \ln m}]$. So, in the initial identity (Theorem 1)

$$|C'| \sum_{j \in I} \binom{m-j}{k-j} B_{2j} > 2^{2k} \frac{m}{m+k} \binom{m+k}{2k},$$

Hence, there exists such a $j \in I$, that

$$B_{2j} > \min_{j \in I} \frac{2^{2k} m \binom{m+k}{2k}}{2c\sqrt{i \ln m}|C'| \binom{m-j}{k-j}(m+k)}.$$

As we noticed in the proof of the previous theorem the minimum is attained at j approximately equal to i, and the claim follows. □

References

[1] P. Delsarte, *An algebraic approach to the association schemes of coding theory*, Philips Research Reports Supplements, No. 10, 1973.

[2] I. Krasikov and S. Litsyn, *Estimates for the range of binomiality in codes' spectra*, IEEE Trans. Inform. Theory, v.43, 3, 1997, pp.987–991.

[3] V. Levenshtein, *Krawtchouk polynomials and universal bounds for codes and designs in Hamming spaces*, v.41, 5, 1995, pp.1303–1321.

[4] J. H. van Lint, *Introduction to Coding Theory*, Springer-Verlag, 1992.

[5] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, New York: North-Holland, 1977.

[6] H. N. Ward, *A bound for divisible codes*, IEEE Trans. Inform.Theory, vol. 38, No. 1, 1992, pp.191–194.

SEMIGROUP RINGS AND THE EXTENSION THEOREM FOR LINEAR CODES

JAY A. WOOD
DEPARTMENT OF MATHEMATICS, COMPUTER SCIENCE & STATISTICS
PURDUE UNIVERSITY CALUMET
HAMMOND, INDIANA 46323–2094

ABSTRACT. An extension theorem for general weight functions is proved over finite commutative local principal ideal rings. The structure of the complex semigroup ring associated to the multiplicative semigroup of the ring plays a prominent role in the proof.

1. BACKGROUND

In her doctoral dissertation, MacWilliams [8], [9] proved an equivalence theorem: two linear codes $C_1, C_2 \subset \mathbb{F}^n$ defined over a finite field $\mathbb{F}$ are equivalent up to monomial transformations if and only if there is a linear isomorphism $f : C_1 \to C_2$ which preserves Hamming weight. Bogart et al. [2] gave another proof of this theorem, and a character theoretic proof was provided by Ward and the author [13].

Following up on the ideas in [13], the author has extended the character theoretic techniques to linear codes defined over finite Frobenius rings, first for the Hamming weight [15] and then for symmetrized weight compositions [16]. In this paper, the author treats general weight functions defined over finite commutative local principal ideal rings. Goldberg proved the extension theorem for symmetrized weight compositions over finite fields, [6], and Constantinescu, Heise, and Honold have proved an extension theorem for homogeneous weight functions over $\mathbb{Z}/m$, [4].

A word on the name of the theorem. MacWilliams' result above is sometimes referred to as "the equivalence theorem of MacWilliams." I have come to prefer "the extension theorem of MacWilliams," because of the similarity to the extension theorems of Witt [14] and Arf [1] for bilinear and quadratic forms. In all these situations there is a fixed ambient space V, usually a finite dimensional vector space over a field. The space V is equipped with an auxiliary function, a weight function in coding theory, a bilinear or quadratic form otherwise. The linear automorphisms of V which preserve the auxiliary function form a group of *linear isometries*, often a classical group in the case of bilinear or quadratic forms, often a group of monomial transformations in coding theory. The *extension theorem* then determines conditions under which any injective linear transformation $f : W \to V$ from a subspace W of V which preserves the auxiliary function must in fact extend to a linear isometry of V itself.

2. STATEMENT OF THE EXTENSION PROBLEM

Fix a finite associative ring R with 1. (Later, we will impose additional hypotheses on R, but we will try to be as general as possible for as long as possible.) Let R^n denote the free module consisting of n-tuples of elements from R. A right *linear code* of length n is a right submodule $C \subset R^n$. The *complete weight composition* is the function

---

Partially supported by NSA grants MDA904-94-H-2025 and MDA904-96-1-0067, and by Purdue University Calumet Scholarly Research Awards.

$n : R \times R^n \to \mathbb{Z}$ given by $n_r(x) = |\{i : x_i = r\}|$, $r \in R$, $x = (x_1, \ldots, x_n) \in R^n$. That is, the complete weight composition counts the number of entries in the n-tuple x which equal a particular element r in R.

Choose complex numbers a_r, $r \neq 0$ in R, and set $a_0 = 0$. Then the *weight function* determined by the a_r's is $w : R^n \to \mathbb{C}$ given by

$$w(x) = \sum_{r \in R} a_r n_r(x), \quad x \in R^n.$$

Since $a_0 = 0$, this sum is the same as the sum over $r \neq 0$.

Example 2.1. For any ring R, choosing $a_r = 1$ for all $r \neq 0$ yields the *Hamming weight.*

Example 2.2. For $R = \mathbb{Z}/k$, thought of as the integers j satisfying $-k/2 < j \leq k/2$, set $a_j = |j|$. The resulting weight function is then the *Lee weight.* The use of the Lee weight for linear codes over $\mathbb{Z}/4$ and its connections with nonlinear binary codes in [3], [7] has been an important motivation for the author's work.

Example 2.3. The *Euclidean weight* for $R = \mathbb{Z}/k$ has $a_j = |j|^2$.

Notice that the Lee and Euclidean weight functions have some symmetry: $a_{-j} = a_j$. More generally, we define the *symmetry group* U of a weight function w by $U = \{u \in \mathcal{U} : a_{ur} = a_r,\ r \in R\}$. Here $\mathcal{U}$ denotes the group of units in R. Since R is finite, all units are necessarily two-sided. The Lee and Euclidean weights have $U = \{\pm 1\}$.

Multiplication defines a left action of the group U on the ring R, with each element acting as an additive group automorphism of R; $u \in U$ defines $r \mapsto ur$. We will write $r \approx s$ if $r = us$ for some $u \in U$. We set $\operatorname{orb}(r) = \{s \in R : r \approx s\}$, the *orbit* of $r \in R$ under U. Of course, $r \approx s$ if and only if $\operatorname{orb}(r) = \operatorname{orb}(s)$. Let $U \backslash R$ be the set of U-orbits in R. Since $a_r = a_s$ if $r \approx s$, the value of a_r depends only on $\operatorname{orb}(r) \in U \backslash R$.

The subgroup U determines a *symmetrized weight composition* swc by

$$\operatorname{swc}_t(x) = |\{i : x_i \approx t\}| = \sum_{r \in \operatorname{orb}(t)} n_r(x).$$

Note that $\operatorname{swc}_s(x) = \operatorname{swc}_t(x)$ if $\operatorname{orb}(s) = \operatorname{orb}(t)$. The weight function w can be written as

$$w(x) = \sum_{t \in U \backslash R} a_t \operatorname{swc}_t(x), \quad x \in R^n. \tag{2.1}$$

Let us now consider the linear automorphisms of R^n which preserve one of our auxiliary functions: either a weight function w or a symmetrized weight composition swc. A right linear transformation $f : R^n \to R^n$ is a right *monomial transformation* if there exist a permutation σ of $\{1, 2, \ldots, n\}$ and units $u_1, u_2, \ldots, u_n$ in R such that

$$f(x_1, x_2, \ldots, x_n) = (u_1 x_{\sigma(1)}, u_2 x_{\sigma(2)}, \ldots, u_n x_{\sigma(n)}),$$

for $(x_1, x_2, \ldots, x_n) \in R^n$. If, in addition, the units $u_1, \ldots, u_n$ lie in a subgroup $U \subset \mathcal{U}$ of the group of units of R, we say that f is a right *U-monomial transformation.* It is easy to verify that the right U-monomial transformations form a group under composition; the group is isomorphic to the n-fold wreath product of U.

In [16], Proposition 2, Theorem 9, the author proved the next two results about automorphisms which preserve a symmetrized weight composition; the second result is the extension theorem for symmetrized weight compositions over finite Frobenius rings. The definition of a Frobenius ring is somewhat technical (see [11] or [15], §2); for finite rings it is equivalent to the character module $\widehat{R} = \operatorname{Hom}_{\mathbb{Z}}(R, \mathbb{T})$ being free as a one-sided R-module. Here, $\mathbb{T}$ is the multiplicative group of unit complex numbers.

Proposition 2.4. *Let $f : R^n \to R^n$ be a right linear automorphism. Then f preserves* swc, *i.e.,* $\mathrm{swc}_t(f(x)) = \mathrm{swc}_t(x)$, *for all $t \in U\backslash R$, $x \in R^n$, if and only if f is a right U-monomial transformation.*

Theorem 2.5. *Suppose R is a finite Frobenius ring, and suppose U is a subgroup of the group of units in R. If $C \subset R^n$ is a right linear code and $f : C \to R^n$ is an injective right linear homomorphism which preserves the symmetrized weight composition* swc, *then f extends to a right U-monomial transformation of R^n.*

The corresponding results for a weight function w are the main items discussed in this paper. For the counterpart of Proposition 2.4, we must assume that the linear automorphism is already a monomial transformation. This result was also proved in [16], Proposition 10.

Proposition 2.6. *Suppose that a weight function w has symmetry group U. Then a right monomial transformation on R^n preserves w if and only if it is a U-monomial transformation.*

The following states the extension problem for weight functions over finite Frobenius rings, the counterpart to Theorem 2.5.

Extension Problem. *Suppose R is a finite Frobenius ring and that w is a weight function with symmetry group U. Determine conditions on the weight function w in order that every injective right linear homomorphism $f : C \to R^n$ which preserves w, $C \subset R^n$ a right linear code, extends to a right monomial transformation on R^n which preserves w, i.e., a right U-monomial transformation on R^n.*

Corollary 2.7. *Suppose w is a weight function with symmetry group U for which the extension problem is solvable. Then the group of right linear automorphisms of R^n which preserve w is exactly the group of right U-monomial transformations.*

Proof. Simply take $C = R^n$. □

3. REDUCING TO THE WEIGHT COMPOSITION CASE

In this section we describe an approach to solving the extension problem for weight functions. This approach reduces the extension problem to the weight composition case covered by Theorem 2.5.

Suppose that $C \subset R^n$ is a right linear code with $f : C \to R^n$ a right linear homomorphism which preserves a weight function w. We assume that w has symmetry group U.

We utilize linearity: both the code C as well as $f : C \to R^n$ are right linear. What happens if we replace x by xs? For convenience, set $\delta_r = \delta_r(x) = n_r(f(x)) - n_r(x)$. Weight preservation says that

$$\sum_{r \neq 0} a_r \delta_r(x) = 0, \quad x \in C.$$

Notice that

$$n_r(xs) = \sum_{q : qs = r} n_q(x).$$

The linearity of C and f allows us to write the weight preservation condition as

$$\sum_{r \neq 0} a_r \left(\sum_{q : qs = r} \delta_q(x) \right) = 0, \quad x \in C, \quad s \in R. \tag{3.1}$$

We write (3.1) in matrix form. Let A be a row vector of length $|R|-1$, with typical entry a_r indexed by $r \neq 0$ in R. Similarly, let Δ be a square matrix of size $|R|-1$ whose rows and columns are indexed by $r, s \neq 0$ in R, with (r,s)-entry equal to $\sum_{q:qs=r} \delta_q(x)$. Thus (3.1) takes the form

$$A\Delta = 0. \tag{3.2}$$

The basic strategy is to exploit the symmetry of w under the group U in order to conclude that $\sum_{q \in \mathrm{orb}(t)} \delta_q(x) = 0$. Since the latter is equivalent to $\mathrm{swc}_t(f(x)) = \mathrm{swc}_t(x)$, swc is preserved by f, and we may apply Theorem 2.5 to prove the extension theorem. How to go about exploiting this symmetry is the subject of subsequent sections.

4. SEMIGROUP RINGS

Let S be a finite semigroup whose operation is written as multiplication. Assume S has both a 0 and a 1 (different). The *complex semigroup ring* $\mathbb{C}[S]$ is then

$$\mathbb{C}[S] = \left\{ \sum_{s \in S} b_s e_s : b_s \in \mathbb{C} \right\},$$

a complex vector space with basis e_s, $s \in S$, whose multiplication is determined by the semigroup multiplication: $e_s e_t = e_{st}$. The one-dimensional subspace spanned by e_0 is a two-sided ideal in $\mathbb{C}[S]$, and its quotient $\mathbb{C}_0[S] = \mathbb{C}[S]/(e_0)$ is called the *reduced complex semigroup ring* associated to S; $\mathbb{C}_0[S]$ has dimension $|S|-1$ over $\mathbb{C}$.

Of particular interest to us is the semigroup we will denote by $S = R^\star$, the multiplicative semigroup of a finite ring. Then $\mathbb{C}_0[R^\star]$ has dimension $|R|-1$. We are interested in the left regular representation of $\mathbb{C}_0[R^\star]$, i.e., $\mathbb{C}_0[R^\star]$ acting on itself by left multiplication. If $b = \sum_{r \neq 0} b_r e_r \in \mathbb{C}_0[R^\star]$, what is the matrix L_b with respect to the basis $\{e_r\}$ for the linear transformation given by left multiplication by b?

The (r,s)-entry of L_b is the coefficient of e_r in be_s. But

$$be_s = \left(\sum_{q \neq 0} b_q e_q \right) e_s = \sum_{r \neq 0} \left(\sum_{q:qs=r} b_q \right) e_r,$$

so that the (r,s)-entry of L_b is $\sum_{q:qs=r} b_q$. This proves the next proposition.

Proposition 4.1. *In (3.2), i.e., $A\Delta = 0$, which expresses the weight preservation property of a linear homomorphism $f : C \to R^n$, the matrix Δ equals the matrix associated to left multiplication by $\delta = \sum_{r \neq 0} \delta_r(x) e_r$ in the reduced complex semigroup ring $\mathbb{C}_0[R^\star]$.*

To exploit the equation $A\Delta = 0$, we seek a better basis for $\mathbb{C}_0[R^\star]$. The basic idea is that the reduced semigroup ring $B = \mathbb{C}_0[R^\star]$ is an Artinian algebra over $\mathbb{C}$, whereby it admits a direct sum decomposition, as a left module over itself,

$${}_B B = \bigoplus V_i, \tag{4.1}$$

where the V_i are indecomposable left B-modules. Each projective indecomposable V_i in turn admits a composition series whose successive quotients are irreducible left B-modules.

By choosing a vector space basis for B over $\mathbb{C}$ which is adapted to the direct sum decomposition and the composition series, the matrix for the left regular representation will take on a block triangular form. If P is the change of basis matrix, expressing the new adapted basis in terms of the old basis of the e_r's, then $P^{-1}\Delta P$ is the matrix for the left regular representation in terms of the new adapted basis; $P^{-1}\Delta P$ is block triangular. Finally, our weight preservation equation, $A\Delta = 0$, can then be written as

$$(AP)(P^{-1}\Delta P) = 0.$$

This decomposition of $\mathbb{C}_0[R^\star]$ can be made very explicit for certain rings, those which are finite commutative local principal ideal rings, as we shall see in Section 6. The decomposition of $\mathbb{C}_0[R^\star]$ will be expressed in terms of the Fourier transform, to which we turn next.

5. FOURIER TRANSFORM

Since we will make heavy use of the Fourier transform for finite abelian groups, this section establishes notation and records some basic facts. All proofs are left to the reader.

Let G be a finite abelian group. A *character* on G is a group homomorphism $\pi : G \to \mathbb{T}$ from G to the multiplicative group of unit complex numbers. The collection of all characters on G, denoted $\widehat{G}$, is itself a finite abelian group under pointwise multiplication. The inverse π^{-1} of π in $\widehat{G}$ is just the complex conjugate $\bar{\pi}$.

Let $f : G \to \mathbb{C}$ be any complex-valued function on G. The *Fourier transform* of f is $\hat{f} : \widehat{G} \to \mathbb{C}$ given by $\hat{f}(\pi) = \sum_{x\in G} f(x)\pi(x)$. This can be inverted.

Proposition 5.1 (Fourier Inversion Formula). *For every* $x \in G$,

$$f(x) = \frac{1}{|G|}\sum_{\pi\in\widehat{G}} \hat{f}(\pi)\pi(x^{-1}) = \frac{1}{|G|}\sum_{\pi\in\widehat{G}} \hat{f}(\pi)\bar{\pi}(x).$$

We want a more general inversion formula, the Poisson summation formula, which depends upon a choice of subgroup $H \subset G$. Define the *annihilator* $(\widehat{G} : H)$ of H in $\widehat{G}$ to be

$$(\widehat{G} : H) = \{\pi \in \widehat{G} : \pi(h) = 1,\ h \in H\}.$$

Then $(\widehat{G} : H) \cong \widehat{G/H}$ and $|(\widehat{G} : H)| = |G|/|H|$.

Lemma 5.2. *For* H *a subgroup of* G *and* $\pi \in \widehat{G}$,

$$\sum_{h\in H} \pi(h) = \begin{cases} |H|, & \pi \in (\widehat{G} : H), \\ 0, & \pi \notin (\widehat{G} : H). \end{cases}$$

Theorem 5.3 (Poisson Summation Formula). *For every* $x \in G$,

$$\sum_{h\in H} f(hx) = \frac{1}{|(\widehat{G} : H)|}\sum_{\pi\in(\widehat{G}:H)} \hat{f}(\pi)\bar{\pi}(x).$$

6. DECOMPOSING THE SEMIGROUP RING

Starting in this section, we concentrate on the case where R is a finite commutative local ring with principal maximal ideal $\mathfrak{m} = Rm$. Our main task is to understand the structure of the reduced semigroup ring $\mathbb{C}_0[R^\star]$. What makes this particularly amenable to attack is the relatively simple ideal structure of the ring R. This and other useful facts are summarized in the next result. The reader may supply the proof or refer to Lemma 13 of [16]. For more detailed results about these rings, see Chapter XVII of [10].

Lemma 6.1. *Let* $(R, \mathfrak{m})$ *be a commutative local ring. Suppose* $\mathfrak{m}$ *is a principal ideal,* $\mathfrak{m} = Rm$. *Then the following hold.*

(i). *There exists an* $l \geq 0$ *such that* $\mathfrak{m}^l \neq 0$, *but* $\mathfrak{m}^{l+j} = 0$ *for all* $j \geq 1$.

(ii). *Each ideal* $\mathfrak{m}^i$ *is principal with* $\mathfrak{m}^i = R(m^i)$.

(iii). *Every ideal in R is equal to one of the $\mathfrak{m}^i$, $i = 0, \dots, l, l+1$.*
(iv). *R is Frobenius.*

Examples of these rings include all finite fields, the rings $\mathbb{Z}/p^{l+1}$ where p is prime, and the Galois rings $GR(p^{l+1}, n)$, [10], Chapter XVI.

The full group of units $\mathcal{U}$ acts on R on the left by multiplication. The orbits are $\mathcal{U} = R \setminus \mathfrak{m}$ itself, $\mathfrak{m} \setminus \mathfrak{m}^2, \mathfrak{m}^2 \setminus \mathfrak{m}^3, \dots, \mathfrak{m}^l \setminus 0$, and 0, where $\setminus$ denotes the set-theoretic difference. Set $\mathcal{O}_0 = \mathcal{U} = R \setminus \mathfrak{m}$, $\mathcal{O}_i = \mathfrak{m}^i \setminus \mathfrak{m}^{i-1}$, and $\mathcal{O}_{l+1} = 0$. Also, set $M_i = |\mathcal{O}_i|$.

Define a relation $\leq$ on R by $y \leq x$ if $y = ax$ for some $a \in R$. The relation $\leq$ is reflexive and transitive, and $x \leq y$, $y \leq x$ imply that x, y lie in the same $\mathcal{U}$-orbit. Thus we see that $\leq$ induces a well-defined partial ordering on the set of $\mathcal{U}$-orbits of R. This works for any finite ring ([15], §5; beware that $\leq$ there is $\geq$ here). What is special for R is that $\leq$ is actually a total ordering, with

$$0 = \mathcal{O}_{l+1} \leq \mathcal{O}_l \leq \mathcal{O}_{l-1} \leq \cdots \leq \mathcal{O}_1 \leq \mathcal{O}_0.$$

We now write down a new basis for the reduced semigroup ring $B = \mathbb{C}_0[R^\star]$. For a character π of the group of units $\mathcal{U}$ and an orbit $\mathcal{O}_i$, we say that the pair $(\pi, \mathcal{O}_i)$ is *admissible* if the stabilizer subgroup $\mathcal{U}_i = \operatorname{Stab}(\mathcal{O}_i)$ of the orbit $\mathcal{O}_i$ is contained in $\ker \pi$. Because R is commutative, $\mathcal{U}_i$ depends only upon the orbit $\mathcal{O}_i$, not on a particular element in the orbit. Note that $\mathcal{U}_i \subset \mathcal{U}_{i+1}$.

Another way to say that the pair $(\pi, \mathcal{O}_i)$ is admissible is that $\pi \in (\widehat{\mathcal{U}} : \mathcal{U}_i)$. Since $|(\widehat{\mathcal{U}} : \mathcal{U}_i)| = |\mathcal{U}|/|\mathcal{U}_i| = |\mathcal{O}_i|$, we see that the number of admissible pairs equals $|R|$. The zero orbit $\mathcal{O}_{l+1}$ has $\mathcal{U}_{l+1} = \mathcal{U}$, so that the trivial character $\pi = 1$ is the only admissible character for $\mathcal{O}_{l+1} = 0$. In dealing with the reduced semigroup ring $\mathbb{C}_0[R^\star]$, the admissible pair $(\pi = 1, \mathcal{O}_{l+1} = 0)$ will be dropped, leaving $|R| - 1 = \dim \mathbb{C}_0[R^\star]$ other admissible pairs.

For each admissible pair $(\pi, \mathcal{O}_i)$, $\pi \in (\widehat{\mathcal{U}} : \mathcal{U}_i)$, define an *admissible orbit sum* by

$$s(\pi, \mathcal{O}_i) = \frac{1}{M_i} \sum_{u \in \mathcal{U}/\mathcal{U}_i} \pi(u) e_{um^i} \in \mathbb{C}_0[R^\star].$$

Recall that $M_i = |\mathcal{O}_i|$, and note that $\mathcal{O}_i = \{um^i : u \in \mathcal{U}/\mathcal{U}_i\}$. Also, $\pi(u)$ for $u \in \mathcal{U}/\mathcal{U}_i$ is well-defined, since $\pi \in (\widehat{\mathcal{U}} : \mathcal{U}_i)$. Notice that $(\pi, \mathcal{O}_0)$ is admissible for every $\pi \in \widehat{\mathcal{U}}$, since $\mathcal{U}_0 = \{1\}$.

For any character $\pi \in \widehat{\mathcal{U}}$, let i_π be the largest integer j such that $(\pi, \mathcal{O}_j)$ is admissible. That is, $\mathcal{U}_{i_\pi} \subset \ker \pi$, but $\mathcal{U}_j \not\subset \ker \pi$ for $j > i_\pi$.

Proposition 6.2. *For any element $b = \sum_{r \neq 0} b_r e_r$ in $\mathbb{C}_0[R^\star]$, and for any admissible orbit sum $s(\pi, \mathcal{O}_i)$, their product satisfies the formula*

$$b\, s(\pi, \mathcal{O}_i) = \sum_{k=0}^{i_\pi - i} \left(\sum_{u \in \mathcal{U}/\mathcal{U}_k} b_{um^k} \bar{\pi}(u) \right) s(\pi, \mathcal{O}_{i+k}).$$

Proof. Group the terms of $b = \sum_{r \neq 0} b_r e_r$ into sums over orbits:

$$b = \sum_{k=0}^{l} \sum_{u \in \mathcal{U}/\mathcal{U}_k} b_{um^k} e_{um^k}.$$

Now focus on the product

$$\left(\sum_{u \in \mathcal{U}/\mathcal{U}_k} b_{um^k} e_{um^k} \right) s(\pi, \mathcal{O}_i) = \frac{1}{M_i} \sum_{u \in \mathcal{U}/\mathcal{U}_k} \sum_{v \in \mathcal{U}/\mathcal{U}_i} b_{um^k} \pi(v) e_{uvm^{i+k}}. \tag{6.1}$$

We wish to know the coefficient of $e_{wm^{i+k}}$, where $w \in \mathcal{U}/\mathcal{U}_{i+k}$. This term will arise from $u \in \mathcal{U}/\mathcal{U}_k$, $v \in \mathcal{U}/\mathcal{U}_i$, which satisfy $uv = w$ in $\mathcal{U}/\mathcal{U}_{i+k}$.

For fixed $w \in \mathcal{U}/\mathcal{U}_{i+k}$ and fixed $u \in \mathcal{U}/\mathcal{U}_k$, select one solution $v_0 \in \mathcal{U}/\mathcal{U}_i$ so that $uv_0 = w$. The other solutions $v = v_0x$ for $uv = w$ are parameterized by elements x in the kernel $\mathcal{U}_{i+k}/\mathcal{U}_i$ of the natural surjection $\mathcal{U}/\mathcal{U}_i \to \mathcal{U}/\mathcal{U}_{i+k}$. In (6.1) there will then be a term of the form

$$\sum_{x\in\mathcal{U}_{i+k}/\mathcal{U}_i} \pi(v_0x) = \begin{cases} (M_i/M_{i+k})\pi(v_0), & \pi \in (\widehat{\mathcal{U}} : \mathcal{U}_{i+k}),\\ 0, & \pi \notin (\widehat{\mathcal{U}} : \mathcal{U}_{i+k}). \end{cases}$$

When $\pi \in (\widehat{\mathcal{U}} : \mathcal{U}_{i+k})$, the value $\pi(v_0)$ depends only on u, w, in which case $\pi(v_0) = \bar{\pi}(u)\pi(w)$. Then the expression in (6.1) simplifies to

$$\left(\sum_{u\in\mathcal{U}/\mathcal{U}_k} b_{um^k}e_{um^k}\right) s(\pi,\mathcal{O}_i) = \begin{cases} \left(\sum_{u\in\mathcal{U}/\mathcal{U}_k} b_{um^k}\bar{\pi}(u)\right) s(\pi,\mathcal{O}_{i+k}), & \pi \in (\widehat{\mathcal{U}} : \mathcal{U}_{i+k}),\\ 0, & \pi \notin (\widehat{\mathcal{U}} : \mathcal{U}_{i+k}). \end{cases}$$

From this the desired formula follows. □

Theorem 6.3. *The structure of the Artinian ring $\mathbb{C}_0[R^\star]$ has the following features.*

(1) The admissible orbit sums $s(\pi, \mathcal{O}_0)$, $\pi \in \widehat{\mathcal{U}}$, are primitive orthogonal idempotents in $\mathbb{C}_0[R^\star]$ whose sum equals 1.

(2) For each character $\pi \in \widehat{\mathcal{U}}$, the subspace spanned by the admissible orbit sums $s(\pi, \mathcal{O}_i)$ is the projective indecomposable submodule V_π generated by $s(\pi, \mathcal{O}_0)$.

(3) For each character $\pi \in \widehat{\mathcal{U}}$, let i_π be the largest integer such that $(\pi, \mathcal{O}_{i_\pi})$ is admissible. Then $(\pi, \mathcal{O}_i)$ is admissible for $i \le i_\pi$. Moreover, the subspaces V_π^i spanned by $s(\pi, \mathcal{O}_i), s(\pi, \mathcal{O}_{i+1}), \ldots, s(\pi, \mathcal{O}_{i_\pi})$ are submodules of $\mathbb{C}_0[R^\star]$, with $\dim_{\mathbb{C}} V_\pi^i = i_\pi - i + 1$. The submodules V_π^i form a composition series for V_π:

$$V_\pi = V_\pi^0 \supset V_\pi^1 \supset \cdots \supset V_\pi^{i_\pi} \supset 0.$$

(4) The left regular representation for an element $b = \sum_{r\ne 0} b_r e_r \in \mathbb{C}_0[R^\star]$, expressed in terms of the basis of admissible orbit sums, has the block diagonal form

$$\begin{pmatrix} \ddots & & \\ & N_\pi & \\ & & \ddots \end{pmatrix},$$

where each N_π is lower triangular of size $(i_\pi+1)\times(i_\pi+1)$. The matrix N_π has the form

$$N_\pi = \begin{pmatrix} \hat{b}(\bar{\pi},0) & 0 & 0 & \cdots & 0 \\ \hat{b}(\bar{\pi},1) & \hat{b}(\bar{\pi},0) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \hat{b}(\bar{\pi},i_\pi-1) & \hat{b}(\bar{\pi},i_\pi-2) & \cdots & \hat{b}(\bar{\pi},0) & 0 \\ \hat{b}(\bar{\pi},i_\pi) & \hat{b}(\bar{\pi},i_\pi-1) & \cdots & \hat{b}(\bar{\pi},1) & \hat{b}(\bar{\pi},0) \end{pmatrix},$$

where

$$\hat{b}(\bar{\pi},j) = \sum_{u\in\mathcal{U}/\mathcal{U}_j} b_{um^j}\bar{\pi}(u), \quad j = 0,1,\ldots,i_\pi.$$

Proof. With the exception of the idempotents $s(\pi, \mathcal{O}_0)$ being primitive, all these statements follow immediately from the formula in Proposition 6.2 and the orthogonality relations on characters.

Since the idempotents $s(\pi,\mathcal{O}_0)$, $\pi\in\widehat{\mathcal{U}}$, are orthogonal, the Artinian ring $B=\mathbb{C}_0[R^\star]$ splits as a direct sum of rings:

$$B=\bigoplus_{\pi\in\widehat{\mathcal{U}}} Bs(\pi,\mathcal{O}_0).$$

The idempotent $s(\pi,\mathcal{O}_0)$ is primitive if and only if the ring $Bs(\pi,\mathcal{O}_0)$ is local.

Inside $Bs(\pi,\mathcal{O}_0)$ consider the subspace V^1_π; it is clearly an ideal. Every element $x\in Bs(\pi,\mathcal{O}_0)$ has the form $x=x_0s(\pi,\mathcal{O}_0)+x'$, where $x_0\in\mathbb{C}$ and $x'\in V^1_\pi$. Standard arguments show that x is a unit in $Bs(\pi,\mathcal{O}_0)$ if and only if $x_0\neq 0$. Thus x is a unit if and only if $x\notin V^1_\pi$. This implies that $Bs(\pi,\mathcal{O}_0)$ is a local ring with maximal ideal V^1_π, as desired. □

7. EXTENSION THEOREM

We continue to assume that $(R,\mathfrak{m})$ is a finite commutative local ring with $\mathfrak{m}=Rm$ principal. We also assume that w is a weight function of the typical form $w(x)=\sum a_r n_r(x)$. In the next result, please be aware that the subgroup U need *not* be the symmetry group of w.

Theorem 7.1. *Let U be any subgroup of $\mathcal{U}$. Suppose the weight function w satisfies $\hat{a}(\pi,i_\pi)=\sum_{u\in\mathcal{U}/\mathcal{U}_{i_\pi}} a_{um^{i_\pi}}\pi(u)\neq 0$ for all $\pi\in(\widehat{\mathcal{U}}:U)$. Then for any right linear code $C\subset R^n$, every injective right linear homomorphism $f:C\to R^n$ which preserves w extends to a U-monomial transformation on R^n.*

Proof. As in the general developments described above, the approach is to utilize the weight preservation equation $A\Delta=0$ in order to show that the symmetrized weight composition determined by U is preserved. The result will then follow from the extension theorem for weight compositions, Theorem 2.5.

We make use of Theorem 6.3 to pick a better basis for $\mathbb{C}_0[R^\star]$: the basis consisting of the admissible orbit sums $s(\pi,\mathcal{O}_j)$, $\pi\in\widehat{\mathcal{U}}$, $j=0,1,\dots,i_\pi$. Then $A\Delta=0$ implies $(AP)(P^{-1}\Delta P)=0$, where P is the change of basis matrix whose columns are the coefficients of the new basis elements $s(\pi,\mathcal{O}_j)$ in terms of the old basis elements e_r.

The entries of the row vector AP are simply $\hat{a}(\pi,i)/M_i$, where

$$\hat{a}(\pi,i)=\sum_{u\in\mathcal{U}/\mathcal{U}_i} a_{um^i}\pi(u).$$

The matrix $P^{-1}\Delta P$ is block triangular, as in Therorem 6.3.

In the matrix equation $(AP)(P^{-1}\Delta P)=0$ the block parameterized by a fixed character π yields the following system of equations:

$$\begin{aligned}
\frac{1}{M_0}\hat{a}(\pi,0)\hat{\delta}(\bar{\pi},0)+\frac{1}{M_1}\hat{a}(\pi,1)\hat{\delta}(\bar{\pi},1)+\cdots+\frac{1}{M_{i_\pi}}\hat{a}(\pi,i_\pi)\hat{\delta}(\bar{\pi},i_\pi)&=0\\
\frac{1}{M_1}\hat{a}(\pi,1)\hat{\delta}(\bar{\pi},0)+\cdots+\frac{1}{M_{i_\pi}}\hat{a}(\pi,i_\pi)\hat{\delta}(\bar{\pi},i_\pi-1)&=0\\
\vdots&=\vdots\\
\frac{1}{M_{i_\pi}}\hat{a}(\pi,i_\pi)\hat{\delta}(\bar{\pi},0)&=0.
\end{aligned}$$

Since we are assuming that $\hat{a}(\pi,i_\pi)\neq 0$ for all $\pi\in(\widehat{\mathcal{U}}:U)$, we see by induction that $\hat{\delta}(\bar{\pi},j)=0$ for $\pi\in(\widehat{\mathcal{U}}:U)$ and $j=0,1,\dots,i_\pi$. Note that $(\pi,\mathcal{O}_j)$ being admissible for $\pi\in(\widehat{\mathcal{U}}:U)$ means that $\pi\in(\widehat{\mathcal{U}}:\mathcal{U}_j)$ as well, so that $\pi\in(\widehat{\mathcal{U}}:U\cdot\mathcal{U}_j)$.

We now apply the Poisson summation formula, Theorem 5.3. The difference of the symmetrized weight compositions has the form

$$\mathrm{swc}_t(f(x)) - \mathrm{swc}_t(x) = \sum_{r \in \mathrm{orb}(t)} (n_r(f(x)) - n_r(x)) = \sum_{r \in \mathrm{orb}(t)} \delta_r(x).$$

More than just a sum over a U-orbit, it is, in fact, a sum over the coset of t for the subgroup $U/(U \cap \mathcal{U}_t)$ of $\mathcal{U}/\mathcal{U}_t$. (We write $\mathcal{U}_t$ for the stabilizer subgroup of t in $\mathcal{U}$. The $\mathcal{U}$-orbit of t is some $\mathcal{O}_i$. The characters of $\mathcal{U}/\mathcal{U}_t$ are exactly those characters for which $(\pi, \mathcal{O}_i)$ is admissible.) The Poisson summation formula states that coset sums are equal to transform sums over annihilators. Since we know that $\hat{\delta}(\bar{\pi}, j) = 0$ for $\pi \in (\widehat{\mathcal{U}} : U)$ and $j = 0, 1, \ldots, i_\pi$, these sums over annihilators vanish. Thus the symmetrized weight composition is preserved, and the theorem follows from Theorem 2.5. □

Corollary 7.2. *Suppose w is any weight function with $a_r > 0$ for $r \neq 0$. Then any weight preserving linear homomorphism on a subspace extends to a monomial transformation.*

Proof. Taking $U = \mathcal{U}$, we see that the only character in $(\widehat{\mathcal{U}} : U)$ is the trivial character $\pi = 1$. For $\pi = 1$, it is clear that $\hat{a}(\pi, i_\pi) \neq 0$, since it equals a sum of positive a_r's. □

The solution of the extension problem now follows from Theorem 7.1 simply by taking U to be the symmetry group of the weight function w. We record this result next.

Theorem 7.3 (Extension Theorem). *Let U be the symmetry group of the weight function w. If $\hat{a}(\pi, i_\pi) = \sum_{u \in \mathcal{U}/\mathcal{U}_{i_\pi}} a_{um^{i_\pi}} \pi(u) \neq 0$ for all $\pi \in (\widehat{\mathcal{U}} : U)$, then the extension problem is solvable for w.*

Remark 7.4. Suppose we are in the situation of Theorem 7.1, where U is some subgroup of $\mathcal{U}$ and $\hat{a}(\pi, i_\pi) \neq 0$ for all $\pi \in (\widehat{\mathcal{U}} : U)$. Then any $f : C \to R^n$ which preserves w extends to a U-monomial transformation on R^n. Beware that the extension of f need not preserve w on all of R^n (a priori, just on C). But if U is a subgroup of the symmetry group of w, then Proposition 2.6 applies to show that the extension also preserves w on R^n.

8. EXAMPLES

We conclude with some examples which illustrate both the uses and the limitations of the theorems of Section 7.

Example 8.1. Let $R = \mathbb{Z}/2^{l+1}$. This is a finite commutative local ring with principal maximal ideal $\mathfrak{m} = (2)$. If we concentrate on the case where $U = \mathcal{U}$, then only the trivial character $\pi = 1$ arises in $(\widehat{\mathcal{U}} : U)$. For $\pi = 1$, $i_\pi = l$, and $\hat{a}(\pi, i_\pi) = a_{2^l}$. Thus, as long as $a_{2^l} \neq 0$, Theorem 7.1 says that a weight preserving $f : C \to R^n$ extends to a monomial transformation. As in Remark 7.4, the extension may not preserve the weight function on all of R^n.

A similar result holds for $R = \mathbb{Z}/p^{l+1}$ with p an odd prime. The condition on w is that

$$a_{p^l} + a_{2p^l} + \cdots + a_{(p-1)p^l} \neq 0. \tag{8.1}$$

Example 8.2. In [4], Constantinescu, Heise, and Honold prove an extension theorem for what they call *homogeneous* weight functions on $\mathbb{Z}/m$. For the case where $m = p^{l+1}$ is a prime power, it is a direct consequence of the definition that homogeneous weight functions satisfy (8.1). Thus the extension theorem in [4], in the case where $m = p^{l+1}$, also follows from Theorem 7.1.

Example 8.3. Here is a further illustration of Remark 7.4. Let $R = \mathbb{F}_5$ and take $a_i = i$, for $i = 0, 1, 2, 3, 4$. The symmetry group of w is trivial.

In R^2, let C be the vector subspace spanned by the vector $(1, 4)$, so that $C = \{(0,0), (1,4), (2,3), (3,2), (4,1)\}$. Every non-zero vector in C has $w(x) = 5$. Thus the linear transformation $f : C \to R^2$ determined by $f(1,4) = (2,3)$ preserves w. Corollary 7.2 says that f extends to a monomial transformation. In fact, f extends to $2I$, i.e., scalar multiplication by 2. But $2I$ does not preserve w on all of R^2.

Now let us try to apply the extension theorem with $U = \{1\}$, the symmetry group of w. The group of units $\mathcal{U}$ is cyclic of order 4 (we will use 2 as a generator), and every character $\pi \in \widehat{\mathcal{U}}$ is admissible. In particular, consider the character π of order 2 in $\mathcal{U}$: $\pi(2^j) = (-1)^j$, for $j = 0, 1, 2, 3$. We then compute that $\hat{a}(\pi) = a_1 - a_2 + a_4 - a_3 = 0$. Thus the extension theorem does not apply.

It is easy to verify that the group of weight preserving automorphisms on R^2 is precisely the symmetric group Σ_2. Since f is not the restriction of a permutation, f does not extend to a weight preserving automorphism.

REFERENCES

[1] Č. Arf, "Untersuchungen über quadratische Formen in Körpern der Charakteristik 2. I.," *J. Reine Angew. Math.*, **183** (1941), 148–167.

[2] K. Bogart, D. Goldberg, and J. Gordon, "An elementary proof of the MacWilliams theorem on equivalence of codes," *Inform. and Control,* **37** (1978), 19–22.

[3] A. R. Calderbank, A. R. Hammons, Jr., P. V. Kumar, N. J. A. Sloane, and P. Solé, "A linear construction for certain Kerdock and Preparata codes," *Bull. Amer. Math. Soc.,* **29** (1993), 218–222.

[4] I. Constantinescu, W. Heise, and Th. Honold, "Monomial extensions of isometries between codes over $\mathbb{Z}_m$," Proceedings of the Fifth International Workshop on Algebraic and Combinatorial Coding Theory (ACCT '96), Sozopol, Bulgaria: Unicorn, Shumen, 1996, pp. 98–104.

[5] C. W. Curtis and I. Reiner, *Representation theory of finite groups and associative algebras,* New York: Interscience Publishers, 1962.

[6] D. Y. Goldberg, "A generalized weight for linear codes and a Witt-MacWilliams theorem," *J. Combin. Theory Ser. A,* **29** (1980), 363–367.

[7] A. R. Hammons, Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, "The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes," *IEEE Trans. Inform. Theory,* **IT-40** (1994), 301–319.

[8] F. J. MacWilliams, "Error-correcting codes for multiple-level transmission," *Bell System Tech. J.,* **40** (1961), 281–308.

[9] ———, "Combinatorial problems of elementary abelian groups," Ph.D. dissertation, Radcliffe College, Cambridge, Mass., 1962.

[10] B. R. McDonald, *Finite rings with identity,* New York: Marcel Dekker, Inc., 1974.

[11] T. Nakayama, "On Frobeniusean algebras. I.," *Ann. of Math.* (2), **40** (1939), 611–633; "II.," **42** (1941), 1–21.

[12] J.-P. Serre, *Linear representations of finite groups,* GTM 42, New York: Springer-Verlag, 1977.

[13] H. N. Ward and J. A. Wood, "Characters and the equivalence of codes," *J. Combin. Theory Ser. A,* **73** (1996), 348–352.

[14] E. Witt, "Theorie der quadratischen Formen in beliebigen Körpern," *J. Reine Angew. Math.,* **176** (1937), 31–44.

[15] J. A. Wood, "Duality for modules over finite rings and applications to coding theory," submitted, 1996.

[16] ———, "Extension theorems for linear codes over finite rings," in *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes,* T. Mora, H. Mattson, eds., LNCS 1255, Berlin:Springer-Verlag, 1997, pp. 329–340.

E-mail address: `wood@calumet.purdue.edu`

On the Algebraic Structure of 2-Observable Controllable Group Codes

Stefano Calabrò, Matteo Campanella, Giovanni Garbo

Dipartimento di Ingegneria Elettronica - University of Palermo, Italy
Viale delle Scienze - Parco d'Orleans - 90128 Palermo - Italy
E-mail: labcom@mbox.unipa.it

Abstract **This work studies the algebraic properties of groups codes. The structure of the state group of a 2-observable shift-invariant complete code is thoroughly defined. We present a procedure that reduces the construction of such group to a finite sequence of group extensions.**

I. INTRODUCTION

The observation that the best coded modulation schemes employ geometrically uniform constellations [1] has strongly stimulated the study of codes over groups, which are the natural counterpart of codes over fields in the classic theory of error correcting codes.

The investigations on group codes split in two branches: the research concerning their algebraic structure and the structure of the corresponding encoders ([2], [3], [4], [5]), and the study of signal space constellations matched to the output groups of the abstract codes ([6], [7]).

Regarding the first branch, a complete theory, comparable to that available for the case of codes over fields, has not been developed yet. However, the research has taken a great advantage from the algebraic approach to system theory due to Willems ([8], [9], [10]). Indeed the observation that a group code can be regarded as a system in the Willems sense, i.e. as a set of behaviors, has led to some amount of convergence between these two research fields. Some outstanding results are the definition of the canonical state group, the characterization of complete codes by means of canonical trellises [4] and the introduction of the concepts of controllability, observability and minimality [11].

When the code is shift-invariant and complete, it can be uniquely defined by a single canonical trellis section. Therefore the investigation on the structure of the code is reduced to the study of the structure of its canonical trellis section. In this work we give a contribution toward the comprehension of the structural properties of a trellis section for a shift-invariant code. All the proofs are reported in appendix.

II. REPRESENTATION OF A SHIFT-INVARIANT GROUP CODE

It is necessary to describe in some detail the point of view and the language adopted in this work. We consider a trellis section without parallel transitions, because they can be introduced separately through a group extension of the branch group. Hence we may define the trellis section by means of the following objects:

a) the **state group** Σ ;
b) the **branch group** B, which is a subgroup of the direct product $\Sigma \times \Sigma$, such that its pro-

This work was supported by Italian Ministry of University and Scientific and Technological Research (MURST).

$$\begin{tikzcd}
\Sigma \arrow[r, "\alpha"] & \Sigma' \\
X \arrow[u, "\lambda"] \arrow[r, "\mu"] & \Sigma \arrow[u, "\beta"']
\end{tikzcd}$$

Fig.1

$$\begin{tikzcd}
\Sigma \arrow[rr, "\alpha"] & & \Sigma' \\
Y \arrow[u, "in"] \arrow[r, "\varphi"] & Z \arrow[r, "in"] & \Sigma \arrow[u, "\beta"']
\end{tikzcd}$$

Fig.2

jections on the factors are the full group Σ (condition of trimness [4]);

c) the **null output group** N which is defined as the subgroup of B corresponding to null output. Of course, N is a normal subgroup of B.

We describe the branch group by means of a *characteristic equation*:

$$B=\{(\sigma_1,\sigma_2)|\alpha(\sigma_1)=\beta(\sigma_2)\}, \quad (1)$$

where α and β are suitable epimorphisms; their image Σ' can be identified with the state group of the state code [4].

It is useful to represent the group N through a pair of *parametric equations*:

$$N=\{(\lambda(x),\mu(x))|x\in X\}, \quad (2)$$

where $X\xrightarrow{\lambda}\Sigma$ and $X\xrightarrow{\mu}\Sigma$ are homomorphisms and x is a *free parameter* which runs over a suitable group X. From the results of [11] it is clear that the minimality of the trellis requires λ and μ to be injections.

The morphisms introduced above can be combined in the diagram of Fig.1, which is commutative because N is a subgroup of B.

By virtue of the observation that λ is injective we can replace X with its isomorphic image $Y\equiv\lambda(X)$. It is also convenient to introduce $Z\equiv\mu(X)$ and to split μ into the product of the isomorphism $Y\xrightarrow{\varphi}Z$ and the inclusion of Z in Σ. Therefore the diagram becomes as it is illustrated in Fig.2.

The condition $N\trianglelefteq B$ implies that the projections Y and Z of N are normal subgroups of Σ. However, in order to give a necessary and sufficient condition for the normality, we must introduce some further notation. Let $\mathrm{Auto}_\Sigma(Y)$ be the subgroup of the group $\mathrm{Auto}(Y)$ of the automorphisms of Y, constituted by the restrictions to Y of the internal automorphisms of Σ:

$$\mathrm{Auto}_\Sigma(Y)=\{a\in \mathrm{Auto}(Y)|a(y)=\sigma^{-1}y\sigma\equiv y^\sigma \text{ for some } \sigma\in\Sigma\}. \quad (3)$$

In a similar way let us introduce $\mathrm{Auto}_\Sigma(Z)$. For each $a\in\mathrm{Auto}(Y)$ we define the isomorphism $\tilde{\varphi}(a)\in\mathrm{Auto}(Z)$ as follows:

$$\tilde{\varphi}(a)(z)=\varphi[a(\varphi^{-1}(z))]. \quad (4)$$

Furthermore, let us define $\Sigma\xrightarrow{can_1}\mathrm{Auto}_\Sigma(Y)$ and $\Sigma\xrightarrow{can_2}\mathrm{Auto}_\Sigma(Z)$ by means of the identities:

$$\begin{aligned}(\mathrm{can}_1(\sigma))(y)&=y^\sigma\\ (\mathrm{can}_2(\sigma))(z)&=z^\sigma\end{aligned} \quad (5)$$

We can now state the following

$$\begin{tikzcd}
 & & & & \boxed{\Sigma'} \arrow[dlll, "p"] \\
\boxed{\Sigma} \arrow[r, "can_1"'] \arrow[urrrr, "\alpha"] & \boxed{\mathrm{Auto}_\Sigma(Y)} \arrow[dr, "\varphi^*"'] & & & \\
 & & \boxed{\mathrm{Auto}_\Sigma(Z)} & & \\
\boxed{Y} \arrow[uu, "in"] \arrow[r, "\varphi"] & \boxed{Z} \arrow[r, "in"] & \boxed{\Sigma} \arrow[u, "can_2"] \arrow[uuurr, "\beta"'] & &
\end{tikzcd}$$

Fig.3

$$\begin{tikzcd}
\mathrm{Auto}(Y) \arrow[r, "\widetilde{\varphi}"] & \mathrm{Auto}(Z) \\
\mathrm{Auto}_\Sigma(Y) \arrow[u, "in"] \arrow[r, "\varphi^*"] & \mathrm{Auto}_\Sigma(Z) \arrow[u, "in"']
\end{tikzcd}$$

Fig.4

Theorem 1:
N is a normal subgroup of B if and only if:

a) $Y \trianglelefteq \Sigma$, $Z \trianglelefteq \Sigma$;

b) there are an epimorphism $\Sigma' \xrightarrow{p} \mathrm{Auto}_\Sigma(Y)$ and an isomorphism $\mathrm{Auto}_\Sigma(Y) \xrightarrow{\varphi^*} \mathrm{Auto}_\Sigma(Z)$ such that the diagrams in Fig.3 and Fig.4 are commutative.

III. CONTROLLABILITY AND OBSERVABILITY

In this section we express the conditions of finite controllability and minimality in the language introduced above.
The condition of *l*-controllability can be expressed in terms of the iterated state codes. Namely, the code sequences of the initial code uniquely define corresponding sequences of states. The group of such sequences can be considered as a new group code which is called the **state code** of the initial code. Let us call $C^{(0)}$ the initial code, $\Sigma^{(0)}$ its state group and $C^{(1)}$ its state code. If we iterate the process we obtain a sequence of codes $\left(C^{(i)}\right)$. It is known in the literature that $C^{(0)}$ is *l*-controllable if and only if $C^{(l+1)} = (e)$ and $C^{(i)} \neq \{e\}$ for $i \leq l$.

The branch group of the code $C^{(1)}$ can be *parametrically* represented as in Fig.5. From this diagram we may obtain a representation of the same branch group through a characteristic equation as in Fig.6. The group $\Sigma^{(2)}$ is defined up to isomorphisms.
Generally, from the diagram of Fig.7 we obtain the diagram of Fig.8.
The condition of *l*-controllability can then be expressed imposing

$$\begin{tikzcd}
\Sigma^{(1)} & \\
\Sigma=\Sigma^{(0)} \arrow[u, "\beta_0"] \arrow[r, "\alpha_0"] & \Sigma^{(1)}
\end{tikzcd}$$

Fig.5

$$\begin{tikzcd}
\Sigma^{(1)} \arrow[r, "\alpha_1"] & \Sigma^{(2)} \\
 & \Sigma^{(1)} \arrow[u, "\beta_1"']
\end{tikzcd}$$

Fig.6

$$\begin{tikzcd}
\Sigma^{(i-1)} \arrow[r, "\alpha_{i-1}"] & \Sigma^{(i)} \\
 & \Sigma^{(i-1)} \arrow[u, "\beta_{i-1}"']
\end{tikzcd}$$

Fig.7

$$\begin{tikzcd}
\Sigma^{(i)} \arrow[r, "\alpha_i"] & \Sigma^{(i+1)} \\
 & \Sigma^{(i)} \arrow[u, "\beta_i"']
\end{tikzcd}$$

Fig.8

$$\begin{tikzcd}
X = X_1 \arrow[r, "\mu = \mu_1"'] & \Sigma \\
 & X = X_1 \arrow[u, "\lambda = \lambda_1"']
\end{tikzcd}$$

Fig.9

$$\begin{tikzcd}
X_1 & \\
X_2 \arrow[u, "\lambda_2"'] \arrow[r, "\mu_2"] & X_1
\end{tikzcd}$$

Fig.10

$$\begin{aligned} &\Sigma^{(l)} = \{e\} \\ &\Sigma^{(i)} \neq \{e\} \quad (i < l) \end{aligned} \tag{6}$$

The condition of *m*-observability can be formulated in a similar way. Specifically, we start from the lower corner of the diagram in Fig.1 and we consider the subgroup of $X_1 \times X_1$ defined by the condition $\mu_1(x) = \lambda_1(x')$, as illustrated in Fig.9. We can give a *parametric* representation of this group, as in Fig.10, where λ_2 and μ_2 are injections. This construction may be indefinitely iterated in an obvious way.

It can be shown that the initial code is *m*-observable if and only if

$$\begin{aligned} &X^{(m)} = \{e\} \\ &X^{(i)} \neq \{e\} \quad (i < m) \end{aligned} \tag{7}$$

IV. STRUCTURE OF THE TRELLIS SECTION OF A 2-OBSERVABLE CODE

In the literature the necessary and sufficient conditions for a group Σ to be the state group of an *l*-controllable and 1-observable code have been presented [5]. Similar conditions for codes which are not 1-observable are not yet known. The principal aim of this work is to characterize the state group Σ of a 2-observable and controllable code. A constructive procedure for such groups will be presented. Such a procedure may be used to implement the structure of the encoder.

By virtue of conditions in Eq.7, the 2-observability implies that the group X_2 is identical, so that $Y \cap Z = \{e\}$ in Fig.3. We can exploit this circumstance to represent the group Σ as a subgroup of a direct product, which is essential for the construction described below. Specifically, the following theorem holds.

Theorem 2:
The homomorphism

$$\Sigma \xrightarrow{\vartheta} \Sigma/Y \times \Sigma/Z: \vartheta(\sigma) = (\sigma Y, \sigma Z) \tag{8}$$

is injective. If $A_1 \equiv \Sigma/Y$, $A_2 \equiv \Sigma/Z$, $Q \equiv \Sigma/YZ$ and $A_1 \xrightarrow{pr_1} Q$, $A_2 \xrightarrow{pr_2} Q$ are the natural projections of A_1 and A_2 on Q, the isomorphic image $\widetilde{\Sigma} \equiv \vartheta(\Sigma)$ is composed with all the pairs which satisfy the equation $pr_1(a_1) = pr_2(a_2)$.

By virtue of this theorem the problem is reduced to the construction of the groups A_1 and A_2 and of the characteristic equation of Σ, represented in $A_1 \times A_2$. We now describe the con-

$$\begin{tikzcd}
S \arrow[r, "\alpha"'] & S' \\
& S \arrow[u, "\beta"]
\end{tikzcd}$$

Fig.11

$$\begin{tikzcd}
H \arrow[r, "\psi_H"] \arrow[dr, "\kappa"] & \Omega \arrow[d, "p_I"] \\
& \overline{\Omega}
\end{tikzcd}$$

Fig.12

struction.

IVA. CONSTRUCTION OF THE STATE GROUP

We start from the diagram, illustrated in Fig.11, of an arbitrary 1-observable controllable code, whose branch group will be denoted U. We will show in the next subsection how to obtain such a diagram.

The first step consists in choosing a normal subgroup $Y \trianglelefteq S$ of S. We will build the groups A_1 and A_2 defining two suitable group operations on the same support set.

The selection of Y in S defines the exact sequence

$$e \to Y \xrightarrow{in} S \xrightarrow{pr} H \equiv S/Y \to e. \tag{9}$$

If we call $\mathrm{Int}(Y)$ the group of the internal automorphisms of Y, we can introduce a natural epimorphism κ between H and $\overline{\Omega} \equiv \dfrac{\mathrm{Auto}_S(Y)}{\mathrm{Int}(Y)}$ in the following way. For each $h \in H$ we choose a representative s of h. This defines an automorphism of $\mathrm{Auto}_S(Y)$: $s \mapsto y^s$. If we change the representative, we generally obtain a different automorphism, which, however, is the same modulo $\mathrm{Int}(Y)$. We then have

$$H \xrightarrow{\kappa} \frac{\mathrm{Auto}_S(Y)}{\mathrm{Int}(Y)} \tag{10}$$

We now represent S as an extension of Y by means of H. Therefore, we choose a representative in S for each element of H. Let us call τ the choice function:

$$\tau(h) = s \mid s \in h. \tag{11}$$

Each element of S is then represented in a unique fashion as $s = \tau(h)y$. From this representation we obtain the system of factors which describes the group operation of the extension.

First we obtain the mapping ψ_H from H in $\mathrm{Auto}_S(Y)$ as

$$\psi_H(h)(y) = y^{\tau(h)}. \tag{12}$$

Obviously we have

$$p_I \circ \psi_H = \kappa, \tag{13}$$

where p_I is the natural projection of $\Omega = \mathrm{Auto}_S(Y)$ over $\overline{\Omega}$ (Fig.12). The system of factors of the extension is defined as

$$\tau(h)\tau(h') = \tau(hh')\chi_H(h,h'), \tag{14}$$

where the mapping χ_H has values in Y.

The epimorphism κ defines a partition of H in cosets. A fundamental result of the theory of group extensions is that it is possible to choose τ in such a way that ψ_H is constant as long as its argument runs over a fixed coset and the same happens for χ_H up to suitable cocycle $z(h_1 h_2)$; we make such a choice for τ. As a consequence, it is possible to define $\overline{\Omega} \xrightarrow{\psi} \Omega$

$$\begin{tikzcd}
\Sigma \arrow[rr, "\mathrm{can}_1"'] & & \mathrm{Auto}_\Sigma(Y) \\
& & \mathrm{Auto}_\Sigma(Z) \arrow[u, "\varphi^{*-1}"'] \\
Y \arrow[uu, "\mathrm{in}"'] \arrow[r, "\varphi"] & Z \arrow[r, "\mathrm{in}"] & \Sigma \arrow[u, "\mathrm{can}_2"']
\end{tikzcd}$$

Fig.13

$$\begin{tikzcd}
S^{(0)} \arrow[r, "\alpha_0"] & e \\
& S^{(0)} \arrow[u, "\beta_0"']
\end{tikzcd}$$

Fig.14

and $\overline{\Omega}\times\overline{\Omega}\xrightarrow{\chi}Y$ such that $\psi_H=\psi\circ\kappa$ and $\chi_H(h_1,h_2)=z(h_1,h_2)\chi(\kappa(h_1),\kappa(h_2))$.

We define $K_{\overline{\Omega}}\equiv(\kappa\times\kappa)\circ(pr\times pr)(U)$ and we choose an arbitrary epimorphism $\Gamma\xrightarrow{\varepsilon}K_{\overline{\Omega}}$. Let φ_1 and φ_2 be the components of ε associated with the two factors of the direct product $\overline{\Omega}\times\overline{\Omega}$.

We observe that a natural action of $\overline{\Omega}$ on the center $Z(Y)$ of Y may be defined. Namely, we put $\zeta^{\overline{\omega}}=\zeta^{\omega}$ with $\omega\in\overline{\omega}$, $\zeta\in Z(Y)$. Starting from this action we define an action on $Z(Y)$ induced by Γ through φ_1 in the following way:

$$\zeta\mapsto\zeta^{\varphi_1(\gamma)}. \tag{15}$$

This action will be denoted $\langle\varphi_1\rangle$.

Similarly we define an action $\langle\varphi_2\rangle$:

$$\zeta\mapsto\zeta^{\varphi_2(\gamma)}. \tag{16}$$

We now introduce the second order cocycle groups

$$\begin{gathered}\mathrm{Z}^2(\Gamma,\mathrm{Z(Y)},\langle\varphi_1\rangle)\\ \mathrm{Z}^2(\Gamma,\mathrm{Z(Y)},\langle\varphi_2\rangle)\end{gathered} \tag{17}$$

i.e. the cocycle groups of Γ with values in $Z(Y)$ associated with the actions $\langle\varphi_1\rangle$ and $\langle\varphi_2\rangle$. If z_1 is an arbitrary cocycle of $\mathrm{Z}^2(\Gamma,\mathrm{Z(Y)},\langle\varphi_1\rangle)$, we define the group A_1 as the set $\Gamma\times Y$ equipped with the operation:

$$(\gamma,\mathrm{y})*(\gamma',\mathrm{y}')=\left(\gamma\gamma',\mathrm{z}_1(\gamma,\gamma')\chi(\varphi_1(\gamma),\varphi_1(\gamma'))\mathrm{y}^{\psi(\varphi_1(\gamma'))}\mathrm{y}'\right). \tag{18}$$

Similarly we define A_2 as the same set equipped with the operation:

$$(\gamma,\mathrm{y})\bullet(\gamma',\mathrm{y}')=\left(\gamma\gamma',\mathrm{z}_2(\gamma,\gamma')\chi(\varphi_2(\gamma),\varphi_2(\gamma'))\mathrm{y}^{\psi(\varphi_2(\gamma'))}\mathrm{y}'\right) \tag{19}$$

with $\mathrm{z}_2\in\mathrm{Z}^2(\Gamma,\mathrm{Z(Y)},\langle\varphi_2\rangle)$.

Finally we define:

$$\begin{tikzcd}
S^{(0)} \arrow[rr, "\alpha_0"] & & e \\
& B_0 \arrow[ul] \arrow[dr] & \\
S^{(1)} \arrow[uu, "\beta_1"] \arrow[ur, "\xi_1"] \arrow[rr, "\alpha_1"] & & S^{(0)} \arrow[uu, "\beta_0"']
\end{tikzcd}$$

Fig.15

$$\begin{tikzcd}
S^{(1)} \arrow[rr, "\alpha_1"] & & S^{(0)} \\
& B_1 \arrow[ul] \arrow[dr] & \\
& & S^{(1)} \arrow[uu, "\beta_1"']
\end{tikzcd}$$

Fig.16

$$\begin{tikzcd}
S^{(l-1)} \arrow[r, "\alpha_{l-1}"] & S^{(l-2)} \\
& S^{(l-1)} \arrow[u, "\beta_{l-1}"']
\end{tikzcd}$$

Fig.17

$$\begin{aligned}
\Sigma &= \{((\gamma, y), (\gamma, y'))\} \\
Y &= \{((e,e),(e,y))\} \\
Z &= \{((e,y),(e,e))\} \\
\varphi &: ((e,e),(e,y)) \mapsto ((e,y),(e,e))
\end{aligned} \tag{20}$$

Theorem 3:

The previous construction generates a 2-observable controllable group code, through Fig.13.

Theorem 4:

Each state group Σ of a 2-observable controllable group code can be obtained up to isomorphisms by means of the above construction.

IVB. CONSTRUCTION OF THE TRELLIS SECTION FOR A 1-OBSERVABLE CONTROLLABLE CODE

In order to implement the diagram of Fig.11, we present here a procedure which reduces the construction of such a diagram to a finite sequence of group extensions.

We start with an arbitrary group $S^{(0)}$ and we define a branch trellis section of a 1-controllable code through the diagram in Fig.14, where α_0 and β_0 are trivial, so that the trellis section is simply the direct product $B_0 \equiv S^{(0)} \times S^{(0)}$.

We now choose a group N_1 and an extension of N_1 by means of B_0, obtaining a new group $S^{(1)}$. We then define the epimorphisms α_1 and β_1 through the diagram in Fig.15.

We then consider the branch trellis section defined by the diagram in Fig.16. We iterate now the construction by choosing a group N_2 and an extension of N_2 through B_1, defining a new group Σ_2. Further iterating the procedure, because of the results illustrated in section III, the diagram in Fig.17 defines a branch trellis section of an l-controllable code.

V. CONCLUSIONS

In this work the structure of the state group of a 2-observable shift-invariant complete code has been studied. We have presented a procedure that reduces the construction of such group to a finite sequence of group extensions.

Future research developments include the generalization of this procedure to a m-observable controllable shift-invariant code. Furthermore the structural characterization of the trellis section may be reflected on the structure of the output group G. Such a study may be useful to investigate the problem of the matching of the output group to signal constellations.

APPENDIX

Proof of Theorem 1:

Necessity:

Property a) is obvious. So we start considering the isomorphism φ^* in b). Let $(\sigma_1,\sigma_2)\in B$: the condition $N \trianglelefteq B$, implies that

$$\forall z \in Z \;\; \exists \bar{z} \in Z \mid \left(\varphi^{-1}(z),z\right)^{(\sigma_1,\sigma_2)} = \left(\varphi^{-1}(\bar{z}),\bar{z}\right). \tag{21}$$

After some simple calculations, exploiting the arbitrariness of z, we get:

$$\tilde{\varphi}\left(\mathrm{can}_1(\sigma_1)\right) = \mathrm{can}_2(\sigma_2) \quad \forall(\sigma_1,\sigma_2)\in B. \tag{22}$$

From Eq.22 we prove, by double inclusion, the relation:

$$\tilde{\varphi}\left(\mathrm{Auto}_\Sigma(Y)\right) = \mathrm{Auto}_\Sigma(Z) \tag{23}$$

Since $\tilde{\varphi}$ is an isomorphism, the latter implies that a unique isomorphism φ^* exists such that the diagram in Fig.4 is commutative.

We consider now the epimorphism p. If we define

$$\overline{B} = \left\{(\sigma_1,\sigma_2) \mid \tilde{\varphi}\left(\mathrm{can}_1(\sigma_1)\right) = \mathrm{can}_2(\sigma_2)\right\}, \tag{24}$$

from Eq.22 we get $B \subseteq \overline{B}$ and consequently

$$\mathrm{Ker}(\alpha) \subseteq \mathrm{Ker}\left(\varphi^* \circ \mathrm{can}_1\right) = \mathrm{Ker}\left(\mathrm{can}_1\right). \tag{25}$$

Therefore *can*$_1$ uniquely factors as $\mathrm{can}_1 = p \circ \alpha$. Similarly $\mathrm{can}_2 = \bar{p} \circ \beta$.

If $\alpha(\sigma) = \beta(\bar{\sigma}) = \sigma'$, it is $(\sigma,\bar{\sigma}) \in B \subseteq \overline{B}$ and by virtue of Eq.24

$$\varphi^* \circ p(\sigma') = \varphi^* \circ \mathrm{can}_1(\sigma) = \mathrm{can}_2(\bar{\sigma}) = \bar{p}(\sigma') \tag{26}$$

Hence $\varphi^* \circ p = \bar{p}$ and the diagram in Fig.3 is commutative.

Sufficiency:

Let $n = \left(\varphi^{-1}(z),z\right)$ be an arbitrary element of N and let (σ_1,σ_2) be an arbitrary element of B. We must prove that an element $\bar{z} \in Z$ exists such that

$$\left(\varphi^{-1}(z),z\right)^{(\sigma_1,\sigma_2)} = \left(\varphi^{-1}(\bar{z}),\bar{z}\right) \tag{27}$$

Because of a), $z^{\sigma_2} \in Z$. Furthermore, from Fig.3 we have:

$$z^{\sigma_2} = \mathrm{can}_2(\sigma_2)(z) = \varphi^* \circ p \circ \beta(\sigma_2)(z) = \varphi^* \circ p \circ \alpha(\sigma_1)(z) = \varphi^* \circ \mathrm{can}_1(\sigma_1)(z) \tag{28}$$

and from Fig.4 we have:

$$\varphi^* \circ \mathrm{can}_1(\sigma_1)(z) = \tilde{\varphi} \circ \mathrm{can}_1(\sigma_1)(z) = \varphi\left[\left(\varphi^{-1}(z)\right)^{\sigma_1}\right] \tag{29}$$

so that $\bar{z}$ can be identified with z^{σ_2} in Eq.27.

Proof of Theorem 2:

- ϑ is injective, because it is $\mathrm{Ker}\,\vartheta = Y \cap Z = \{e\}$.
- $(a_1,a_2) \in \tilde{\Sigma} \Rightarrow \mathrm{pr}_1(a_1) = \mathrm{pr}_2(a_2)$, because it is $a_1 = \sigma Y$ and $a_2 = \sigma Z$, so that $\mathrm{pr}_1(a_1) = \sigma YZ = \sigma ZY = \mathrm{pr}_2(a_2)$.
- $\mathrm{pr}_1(a_1) = \mathrm{pr}_2(a_2) \Rightarrow (a_1,a_2) \in \tilde{\Sigma}$. Indeed, let $a_1 = \sigma_1 Y$ and $a_2 = \sigma_2 Z$. The condition

$\sigma_1 \mathrm{YZ} = \sigma_2 \mathrm{YZ}$ implies that there are $\mathrm{y} \in \mathrm{Y}$ and $\mathrm{z} \in \mathrm{Z}$ such that $\sigma_2 = \sigma_1 \mathrm{yz}$. If we define $\sigma = \sigma_1 \mathrm{y} = \sigma_2 \mathrm{z}^{-1}$, we get $(\mathrm{a}_1, \mathrm{a}_2) = \vartheta(\sigma)$.

Proof of Theorem 3:

Since $\mathrm{K}_{\overline{\Omega}}$ is obtained from the branch group U of a controllable code collapsing whole classes of states in a single state, it can be regarded as the branch group of a controllable code too.

We also define

$$\mathrm{K}_{\Omega} = (\mathrm{pr} \times \mathrm{pr})^{-1}\left(\mathrm{K}_{\overline{\Omega}}\right) \subseteq \Omega \times \Omega \tag{30}$$

where pr is the natural projection from Ω to $\overline{\Omega}$. It can be shown that K_{Ω} is the branch group of a controllable code and can be expressed as

$$\mathrm{K}_{\Omega} = \left\{\left(\varphi^{*-1} \circ \mathrm{can}_2(\sigma), \mathrm{can}_1(\sigma)\right)\right\} \tag{31}$$

Hence K_{Ω} can be thought as the branch group of the state code (Fig.13). The fact that the state code is controllable implies that the code obtained by this construction is controllable too; indeed the latter can be obtained from K_{Ω} applying a single step of the construction described in Section IVb. The 2-observability stems from $\mathrm{Y} \cap \mathrm{Z} = \{\mathrm{e}\}$.

Proof of Theorem 4:

Let assume that the code defined by Fig.3 is 2-observable and controllable. We can consider a new code whose state group is Σ, whose branch group is $\overline{\mathrm{B}}$ (Eq.24) and whose null output group is defined by $\mathrm{Y} \xrightarrow{\varphi} \mathrm{Z}$ as for the code in Fig.3. Such a code is still 2-observable and controllable because $\mathrm{B} \subseteq \overline{\mathrm{B}}$.

Starting from this code we now define inputs for the construction in order to recover an isomorphic image of Σ

We define:

$$\Gamma = \Sigma / \mathrm{YZ}. \tag{32}$$

Since we may regard Σ / Y as an extension of ZY/Y by means of Γ, we can define a homomorphism j from Γ to $\dfrac{Auto(ZY/Y)}{Int(ZY/Y)}$. We can transfer j into a homomorphism from Γ to $\dfrac{Auto(Y)}{Int(Y)}$ using φ and the natural isomorphism between Z and ZY/Y. The restriction of this transferred homomorphism to its image will be denoted φ_1.

In a similar way, using Σ / Z, we can introduce φ_2:

$$\Gamma \xrightarrow{\varphi_2} \frac{Auto_{\Sigma}(Y)}{Int(Y)} \tag{33}$$

We now define $\varepsilon(\gamma) = (\varphi_1(\gamma), \varphi_2(\gamma))$ and $K_{\overline{\Omega}} = \{\varepsilon(\gamma)\}$. We have $K_{\overline{\Omega}} = (pr \times pr)(K_{\Omega})$ with K_{Ω} given by Eq.31. Let ψ be an arbitrary choice function from $\overline{\Omega}$ in Ω and χ an associated factor system. Let $H \xrightarrow{\kappa} \overline{\Omega}$ be an arbitrary epimorphism.

S is the set $H \times Y$ equipped with the operation:

$$(h, y)(h', y') = \left(hh', \chi(\kappa(h), \kappa(h')) y^{\psi(\kappa(h'))} y'\right) \tag{34}$$

We choose U as the group:

$$U=\left(\kappa\times\kappa\circ pr\times pr\right)^{-1}\left(K_{\overline{\Omega}}\right) \tag{35}$$

where pr is the projection of S over H. U is the branch group of a controllable code.

A presentation of A_1 as an extension of Y by means of Γ can be obtained choosing $\psi_1=\psi\circ\varphi_1$.

A factor system χ_1 for this presentation must satisfy the equation:

$$I\left[\chi_1(\gamma_1,\gamma_2)\right]=\psi_1(\gamma_1,\gamma_2)^{-1}\,\psi_1(\gamma_1)\psi_1(\gamma_2) \tag{36}$$

where $I(x)$ is the internal automorphism associated with x. A similar equation holds for χ:

$$I\left[\chi(\omega_1,\omega_2)\right]=\psi_1(\omega_1,\omega_2)^{-1}\,\psi_1(\omega_1)\psi_1(\omega_2). \tag{37}$$

Substituting $\omega=\varphi_1(\gamma)$ in Eq.37 and comparing Eqs.36 and 37, we get:

$$\chi_1(\gamma_1,\gamma_2)=z_1(\gamma_1,\gamma_2)\chi\left[\varphi_1(\gamma_1),\varphi_1(\gamma_2)\right] \tag{38}$$

where z_1 is a suitable second order cocycle belonging to $Z^2\left[\Gamma,Z(Y),\langle\varphi_1\rangle\right]$. In a similar way we define a cocycle $z_2\in Z^2\left[\Gamma,Z(Y),\langle\varphi_2\rangle\right]$.

We now show that the application of the construction to these data returns a group $\tilde{\Sigma}$ isomorphic to Σ.

By virtue of Eq.32, the projections from A_i $(i=1,2)$ into Γ are implemented as:

$$pr_i(a_i)=pr_i\left[(\gamma,y)\right]=\gamma \quad (i=1,2) \tag{39}$$

Thus, owing to Theorem 2, the group $\tilde{\Sigma}$ is isomorphic to Σ.

REFERENCES

[1] G. D. Forney, Jr., "Geometrically uniform codes", IEEE *Trans. Inform. Theory*, vol.37, no.5, pp.1241-1260, Sept. 1991

[2] H.-A. Loeliger, "On Euclidean-space group codes", Ph.D. dissertation, ETH Zürich, Switzerland, ETH-Diss 9720, 1992

[3] M. D. Trott, "The algebraic structure of trellis codes", Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, 1992

[4] G. D. Forney, Jr. and M. D. Trott, "The dynamics of group codes: state spaces, trellis diagrams and canonical encoders", IEEE *Trans. Inform. Theory*, vol.39, no.5, pp.1491-1513, Sept. 1993

[5] H.-A. Loeliger and T. Mittelholzer, "Convolutional codes over groups", IEEE *Trans. Inform. Theory*, vol.42, no.6, pp.1660-1686, Nov. 1996

[6] S. Benedetto, R. Garello, M. Mondin, and G. Montorsi, "Geometrically uniform partitions of L×MPSK constellations and related binary trellis codes", IEEE *Trans. Inform. Theory*, vol.39, no.6, pp.1773-1798, Nov. 1993

[7] S. Benedetto, R. Garello, M. Mondin, and G. Montorsi, "Geometrically uniform TCM codes over groups based on L×MPSK constellations", IEEE *Trans. Inform. Theory*, vol.40, no.1, pp.137-152, Jan. 1994

[8] J. C. Willems, "From time series to linear systems-Part I: Finite dimensional linear time invariant systems", *Automatica*, vol.22, pp. 561-580, 1986

[9] J. C. Willems, "Models for dynamics", in *Dynamics Reported*, vol.2, U. Kirchgraber and H. O. Walther, Eds. New York: Wiley and Teubner, 1989

[10] J. C. Willems, "Paradigms and puzzles in the theory of dynamical systems", IEEE Trans. Automat. Contr., vol.36, pp. 259-294, Mar. 1991

[11] H.-A. Loeliger, G. D. Forney, Jr., T. Mittelholzer, and M. D. Trott, "Minimality and observability of group systems", *Linear Alg. Appl.*, vol. 205-206, pp. 937-963, July 1994

Multidimensional Convolutional Codes *

Paul A. Weiner
Department of Mathematics
University of Notre Dame

In this short note we consider some aspects of multidimensional convolutional codes, paying special attention to some significant differences between 1-D, 2-D, and higher-D convolutional codes.

Throughout we let $\mathbb{F} = \mathbb{F}_q$ be the finite field of q elements, and $R = \mathbb{F}[z_1, ..., z_m]$ be the polynomial ring in m indeterminates over $\mathbb{F}$.

Definition 1 An *m-dimensional (m-D) convolutional code of length n* is an R-submodule of R^n. Hence any m-D convolutional code can be expressed as rowspace(G), where $G \in R^{l \times n}$; and conversely any matrix $G \in R^{l \times n}$ defines a m-D convolutional code via its rowspace. If a code is the rowspace of the matrix G, we say G is a *generator matrix* for the code. We will write $\mathcal{C}$ for a convolutional code, and $\mathcal{C}(G)$ for the convolutional code with generator matrix G.

A 0-D convolutional code is simply a block code, and a 1-D convolutional code is a convolutional code as defined by York [3]. 2-D convolutional codes (with a slightly different definition) are studied by Fornasini and Valcher [1, 2].

Because a polynomial ring in one variable over a field is a principal ideal domain (and this fails for more than one variable), certain results standard in the theory of 1-D convolutional codes no longer hold, or require modification, in higher dimensions. There are also interesting differences between 2-D codes and higher-D codes.

Definition 2 Let $G \in R^{l \times n}$. The *rank* of the code $\mathcal{C}(G)$ is the largest integer k for which there is a nonzero $k \times k$ minor of G. If $G \in R^{k \times n}$, and $\mathcal{C}(G)$ is of rank k, then we say $\mathcal{C}(G)$ is a *free* convolutional code (it is isomorphic to the free module R^k).

Rank and freeness are properties of a code and are independent of generator matrix chosen for the code. A code is free of rank k if and only if it has a $k \times n$ generator matrix with at least one nonzero $k \times k$ minor. Free codes are advantageous in that a $k \times n$ generator matrix of a rank k free code serves as an encoder via matrix multiplication. Every 1-D code is free, but this is no longer true for 2-D codes.

Example 3 The 2-D code whose generator matrix is given by $G = \begin{bmatrix} z_1 & {z_1}^2 & z_1 z_2 \\ z_2 & z_1 z_2 & {z_2}^2 \end{bmatrix}$ is not free (it is rank 1 but has no 1×3 generator matrix).

Definition 4 Let $G \in R^{k \times n}$ be of rank k. Let $S = \{\text{full-size minors of } G\}$. Let J be the ideal $(S) \subseteq R$. G is
i) *unimodular* if $k = n$ and $\det(G)$ is a unit in R (i.e., a nonzero constant);
ii) *zero prime* (ZP) if $J = R$;
iii) *minor prime* (MP) if the elements of S have no nonconstant common factor;
iv) *left-factor prime* (LFP) if any factorization $G = T \cdot \bar{G}$ with $T \in R^{k \times k}$ and $\bar{G} \in R^{k \times n}$ has T unimodular.

---

*This work was supported in part by a fellowship from The Center for Applied Mathematics, University of Notre Dame.

Proposition 5 *See [4]. With notation as in the previous definition,*
i) ZP $\Rightarrow$ MP $\Rightarrow$ LFP.
ii) If $m = 1$, then all are equivalent.
iii) If $m = 2$, then MP $\Leftrightarrow$ LFP (and ZP is different).
iv) If $m > 2$, then all are different.

Zerz [4] gives the important example over the complex numbers of a matrix that is LFP but not MP. We note that this example works over general fields by the same reasoning given by Zerz. We give the example in the binary case.

Example 6 Over $\mathbb{F}_2$, the matrix $G = \begin{bmatrix} z_1 & 0 & z_2 \\ 0 & z_1 & z_3 \end{bmatrix}$ is LFP but not MP.

Definition 7 Let $G \in R^{k \times n}$ generate a free code $\mathcal{C}$ of rank k. A *parity check matrix* for $\mathcal{C}$ is a matrix $H \in R^{l \times n}$ such that $\mathcal{C} = \{w \in R^{1 \times n} : wH^t = 0\}$.

In the 1-D case it is well known that a code has a parity check matrix if and only if the code has a ZP generator matrix. In the general case the following is true.

Proposition 8 *Let $G \in R^{k \times n}$ generate a free code $\mathcal{C}$ of rank k. $\mathcal{C}$ has a parity check matrix if and only if G is minor prime.*

In the block code case the "dual code" of $\mathcal{C}$ is the code generated by the rows of a parity check matrix H of $\mathcal{C}$. The same can be done for an m-D code, but there is a difficulty: the dual code need not be free.

Example 9 Let $\mathcal{C}$ be the 3-D free rank 1 code over $\mathbb{F}_2$ with generator matrix

$$G = \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix}$$

A parity check matrix for $\mathcal{C}$ is

$$H = \begin{bmatrix} z_2 & z_1 & 0 \\ z_3 & 0 & z_1 \\ 0 & z_3 & z_2 \end{bmatrix}$$

The dual code $\mathcal{C}(H)$ is of rank 2 and not free.

Note that it can be shown that examples of this type cannot occur in dimension 2 or lower.

References

[1] E. Fornasini and M.E. Valcher. Algebraic aspects of 2D convolutional codes. *IEEE Trans. Inform. Theory*, IT-40(4):1068–1082, 1994.

[2] M.E. Valcher and E. Fornasini. On 2D finite support convolutional codes: an algebraic approach. *Multidim. Sys. and Sign. Proc.*, 5:231–243, 1994.

[3] E.V. York. *Algebraic Description and Construction of Error Correcting Codes, a Systems Theory Point of View.* PhD thesis, University of Notre Dame, 1997. Available at http://www.nd.edu/~rosen/preprints.html.

[4] E. Zerz. Primeness of multivariate polynomial matrices. *Systems & Control Letters*, 29:139–145, 1996.

Recursive Maximum Likelihood Decoding Algorithm for Convolutional Codes and Its Complexity

Hiroshi Yamamoto Toru Fujiwara Takayuki Uchida

Abstract

A recursive maximum likelihood decoding algorithm (RMLD) is applied to binary convolutional codes. The computational complexity of the RMLD algorithm is analyzed in terms of the number of addition-equivalent operations.

1 Introduction

For linear block codes, it is shown in [1] that the complexity of maximum likelihood decoding (MLD) can be reduced considerably by using the Viterbi decoding based on the optimally sectionalized trellis diagram, where the computational complexity of the algorithm is analyzed in terms of the number of addition-equivalent operations. A trellis-based MLD Algorithm that is more effective than the Viterbi algorithm has recently been proposed [2, 3]. This new algorithm is called RMLD because it is recursive in nature.

Convolutional codes are linear codes with infinite code length. In this paper, we apply RMLD algorithm to binary convolutional codes. For some example codes, it is shown that we can decode the convolutional codes with much less complexity.

2 RMLD Algorithm for Convolutional Code

For a binary convolutional code C, Let T_C denote the minimal trellis of C. For time x and y with $0 \leq x < y$, let $L_{x,y}$ denote the family of the sets of paths of T_C from states in time x to states in time y. For $D \in L_{x,y}$, let $l(D)$ and $m(D)$ denote the path in D which has the largest metric in D and its metric, respectively. A table consisting of $l(D)$ and $m(D)$ for each $D \in L_{x,y}$ is called $L_{x,y}$-table. When $L_{0,t}$-table is constructed, it stores the most likely path for the received symbols from time 1 to time t and its metric for each state in time t among the paths from the initial state to the state. The RMLD algorithm for constructing $L_{0,t}$-table is simply a recursive procedure for constructing $L_{x,y}$-tables ($0 \leq x < y \leq t$) for long trellis sections from short trellis sections. In [3], two algorithms to perform this are shown. The simpler one is called ACS (add-compare-select) and more effective one is called SACS which uses detailed structure of trellis diagram. In both algorithms, we need to choose sectionalization (the way of dividing the problem recursively) properly to obtain a decoder with the small number of operations. A sectionalization which gives the smallest decoding complexity is called the optimum sectionalization for the code. We can get the optimum sectionalization in dynamic programming manner efficiently.

To avoid large decoding delay, we can restrict the sectionalization to the following serially decodable sectionalization with

---

Dept. of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560 Japan
TEL: +81-6-850-6563 FAX: +81-6-850-6564
E-mail: hiroshi@ics.es.osaka-u.ac.jp
(Hiroshi Yamamoto)

maximum section length c; Where c is a relatively small constant: $L_{x,y}$-table is constructed from $L_{x,z}$- and $L_{z,y}$-table with $y - c \leq z < y$. In general, optimum sectionalization is not necessarily serially decodable one with small maximum section length. In this case, we can find the sectionalization which gives the smallest computational complexity among serially decodable sectionalization with maximum section length c for a given integer c.

3 Examples

We compute the optimum sectionalizations as well as the complexities for convolutional codes with code rate $R = 2/3$ and $K = 9, 10$ which is presented in Section 11.3 of [4] as the codes whose minimum free distances are maximum.

Both codes have serially decodable optimum sectionalizations and the maximum section length $c = 6$ is sufficient to attain an optimum sectionalization. Moreover, the sectionalizations are serially decodable with maximum section length $c = 3$ except for the construction of the $L_{0,18}$-table.

Figure 1 shows the computational complexity of the convolutional code with $K = 10$. The number of operations required to construct $L_{0,t}$-table is plotted as a function of time t. The complexity of conventional Viterbi decoding, that of RMLD using ACS, and that of RMLD using SACS are denoted "Vitebi," "ACS," and "SACS" respectively. The complexity of RMLD using SACS is evaluated by an upper bound which is shown in [3] because it depends on channel noise. For both codes, RMLD using SACS gives the smallest complexity which is 48% of the complexity of conventional Viterbi decoding.

The fact that the decoding complexity grows linearly for sufficiently large t suggests that an optimum sectionalization has simple iterative structure. In fact, an optimum sectionalization of each example codes has simple iterative structure for large t.

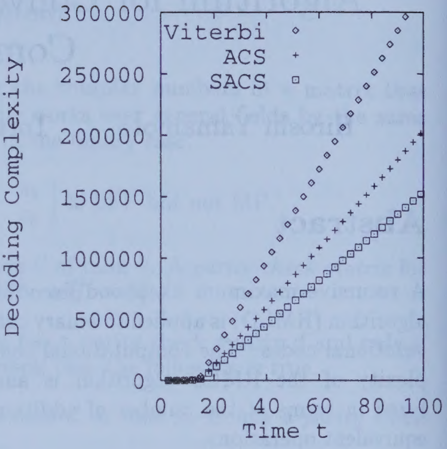


Figure 1: Decoding complexity for ACS of the example code with $R = 2/3$, $K = 10$

References

[1] A. Lafourcade and A. Vardy, "Optimum Sectionalization of a Trellis," IEEE Trans. Inf. Theory, Vol.42, No.3, pp.689–702, May 1996.

[2] H. Yamamoto, H. Nagano, T. Fujiwara, T. Kasami, and Shu Lin, "Recursive MLD Algorithm Using the Detail Trellis Structure for a Linear Block Code and Its Average Complexity Analysis," 1996 International Symposium on Information Theory and Its Applications, pp.704–708, Sep. 1996.

[3] T. Fujiwara, H. Yamamoto, T. Kasami, and S. Lin, "A Trellis-Based Recursive Maximum Likelihood Decoding Algorithm for Linear Block Codes," to appear in IEEE Trans. Inf. Theory, 1998.

[4] S. Lin and D.J. Costello Jr., "Error Control Coding: Fundamentals and Applications," Prentice-Hall, 1983.

Numerical Analysis of the Maximum A Posteriori Algorithm

Mark C. Reed, Paul D. Alexander and John A. Asenstorfer
Mobile Communications Research Centre
Institute for Telecommunications Research
University of South Australia
SPRI Building, Warrendi Road
The Levels SA 5095, Australia
FAX: +61 8 8302 3873

Abstract

This paper describes an analytical method of computing the performance improvement of the Maximum A Posteriori (MAP) decoder. The analysis takes expectations of the output conditional probabilities and manipulates this result to produce an output noise variance. Comparisons to a simulated MAP decoder are shown with close compliance between the two methods.

1 Introduction

Recently a new coding method, called Turbo Codes, was introduced [1]. This technique achieves reliable transmission while operating close to the Shannon limit. Turbo codes combine the concept of soft-in/soft-out decoding, iterative decoding, non-uniform random interleaving, and parallel concatenated convolutional codes (PCCC). Analysis of this technique has been published by Benedetto *et al.* [2,3] and Perez *et al.* [4] but most discuss bounds to turbo codes.

The main decoding component of turbo codes is a soft-in/soft-out convolutional decoder. A Maximum A Posteriori (MAP) decoder [5] provides the best known decoding performance in turbo decoders. The MAP algorithm is useful as it minimises the symbol error probability, and as well produces soft output information. The MAP decoder was originally developed by Bahl *et al.* [5] as an alternative to the Viterbi Algorithm [6] and uses the familiar trellis structure as first described by Viterbi [6]. We are motivated to analyse the performance of turbo codes, and more generally iterative systems based on this structure, such as the iterative multiuser receiver [7]. We therefore need to analyse the operation of the MAP decoder.

We have chosen to view the MAP decoder as a noise variance reduction device and therefore desire to describe the output noise variance of the decoder in terms of its input noise variance. In [8] an estimate of this relationship is given for a small range of signal to noise ratios based on simulation results. In this paper we firstly describe the operation

---

This work was supported in part by Telstra Australia under Contract No. 700136 and by the Commonwealth of Australia under International S & T Grant No. 56.

of the MAP decoder, we then analyse the decoder and derive a method of numerically determining the output variance in terms of the input variance. We then compare a simulated MAP decoder result with the numerical result and show they are very similar. Future work will use the results presented in this paper to describe the performance of turbo codes and iterative systems in general.

Throughout this paper scalars are lower case and vectors are underlined lower case.

2 Channel Model

We model the channel as a coded, discrete-time system, with binary phase shift keying (BPSK) modulation. The channel adds zero-mean white Gaussian noise with variance $\sigma_i^2 = N_0/2$, where N_0 is the single sided noise power spectral density. We assume perfect symbol synchronisation with a block of L coded symbols where $d_{t,n}$ identifies the coded bit n at time t, where $t \in \{0, \cdots, T-1\}$ identifies the information symbol interval, and $n \in \{0, \cdots, r-1\}$, for a rate $R = 1/r$ code, identifies the coded bit n at this time t. The input to the decoder is then

$$\underline{y} = \underline{d} + \underline{n}$$

where $d_{t,n} \in \{-1, +1\}$. For the purposes of analysis we assume that the data transmitted is the all zero path.

We also want to model the output of the MAP decoder. We assume that the output of the MAP decoder is

$$\underline{y}' = \underline{d} + \underline{n}'$$

Where $n'_{t,n} \sim N(0, \sigma_o^2)$.

3 MAP Decoder

As the MAP decoder forms a central component in this paper we will briefly discuss it's operation as documented in [5, 9]. The MAP decoder's purpose is to estimate the individual a-posteriori channel symbol probabilities $\Pr\{d_{t,n}|\underline{y}\}$ or the a-posteriori information bit probabilities $\Pr\{b_t|\underline{y}\}$, for a given received sequence $\underline{y}$. In the special case of a convolutional encoder the MAP achieves this by traversing the decoding trellis in a backward and forward direction; if S_{t-1} and S_t are two states of the code trellis at time $t-1$ and t the MAP decoder calculates

$$\begin{aligned}\Pr\{S_{t-1} = m'; S_t = m|\underline{y}\} &= \Pr\{S_{t-1} = m'; b_t = b|\underline{y}\} \\ &= \Pr\{S_{t-1} = m'; \underline{d}_t = \underline{d}|\underline{y}\} \qquad (1)\end{aligned}$$

where b_t is the information bit which causes the transition $m' \rightarrow m$, and $\underline{d}_t$ are the coded symbols on that transition. These expressions are used to determine

$$\Pr\{d_{t,n} = d|\underline{y}\} = \sum_{m'} \sum_{\substack{\underline{d} \\ (d_{t,n}=d)}} \Pr\{S_{t-1} = m'; \underline{d}_t = \underline{d}|\underline{y}\} \qquad (2)$$

and

$$\Pr\{b_t|\underline{y}\} = \sum_{m'} \Pr\{S_{t-1} = m'; b_t = b|\underline{y}\}$$

where $d \in \{-1, +1\}$. The MAP decoder calculates (1) by evaluating the joint probabilities (applying Bayes' theorem)

$$\Pr\{S_{t-1} = m', S_t = m|\underline{y}\} = \frac{\Pr\{S_{t-1} = m', S_t = m, \underline{y}\}}{\Pr\{\underline{y}\}} \tag{3}$$

This joint probability can be expressed as [5, 9]

$$\Pr\{S_{t-1} = m', S_t = m, \underline{y}\} = \beta_t(m)\gamma_t(m', m)\alpha_{t-1}(m')$$

where the reverse state probability is defined recursively as

$$\begin{aligned} \beta_t(m) &= \Pr\{\underline{y}_{t+1}^T | S_t = m\} \\ &= \sum_{m'} \beta_{t+1}(m')\gamma_{t+1}(m, m') \end{aligned}$$

and the forward state probability is also defined recursively as

$$\begin{aligned} \alpha_t(m) &= \Pr\{S_t = m; \underline{y}_1^t\}, \\ &= \sum_{m'} \alpha_{t-1}(m')\gamma_t(m', m) \end{aligned}$$

where $\underline{y}_{t+1}^T = (y_{t+1}, \ldots, y_T)^\intercal$, and $y_t = \{y_{t,1}, \ldots, y_{t,r}\}$. The transition metric, $\gamma_t(m', m)$, is equal to

$$\begin{aligned} \gamma_t(m', m) &= \Pr\{S_t = m; y_t | S_{t-1} = m'\} \\ &= \sum_{b=b_t} p(y_t|b)\Pr\{b|m, m'\}\Pr\{S_t = m|S_{t-1} = m'\} \end{aligned} \tag{4}$$

where $\Pr\{S_t = m|S_{t-1} = m'\}$ is the state transition probability, or the a-priori transition probability. The distribution $p(y_t|b)$ is the memoryless channel symbol distribution given the transmitted information bit. When the transition from state m to m' exists for a known value of b, and there is only one transition then $\Pr\{b|m, m'\} = 1$, and (4) can be simplified to

$$\gamma_t(m', m) = p(y_t|b)\Pr\{S_t = m|S_{t-1} = m'\}$$

this can also be written in terms of the coded bits as

$$\gamma_t(m', m) = \Pr\{S_t = m|S_{t-1} = m'\} \prod_{n=1}^{r} p(y_{t,n}|d_{t,n})$$

If there is no a-priori information, as assumed in this paper, then $\Pr\{S_t = m|S_{t-1} = m'\} = \frac{1}{2}$ if the transition exists.

4 Analysis of the MAP Decoder

We are interested in the conditional probability of the coded output symbols, which is a step towards the noise perturbed signal, therefore we restate (2) as

$$\begin{aligned}\Pr\{d_{t,n} = d|\underline{y}\} &= \sum_{m'} \sum_{\substack{\underline{d}\\(d_{t,n}=d)}} \Pr\{S_{t-1} = m'; \underline{d}_t = \underline{d}|\underline{y}\} \\ &= \sum_{m'} \sum_{\substack{\underline{d}\\(d_{t,n}=d)}} \beta_t(m)\gamma_t(m',m)\alpha_{t-1}(m')\end{aligned}$$

where the term $\Pr\{\underline{y}\}$ in (3) is a constant for all transitions and can be ignored. We desire the average result of the above, as a step in computing the output noise variance, so taking expectations we get

$$\begin{aligned}\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n} = d|\underline{y}\}\} &= \mathrm{E}_{\underline{n}}\left\{\sum_{m'} \sum_{\substack{\underline{d}\\(d_{t,n}=d)}} \beta_t(m)\gamma_t(m',m)\alpha_{t-1}(m')\right\} \\ &= \sum_{m'} \sum_{\substack{\underline{d}\\(d_{t,n}=d)}} \mathrm{E}_{\underline{n}}\{\beta_t(m)\gamma_t(m',m)\alpha_{t-1}(m')\} \qquad (5)\end{aligned}$$

As β_t, γ_t and α_{t-1} are independent of each other and due to the independent and identically distributed (i.i.d.) data and noise samples we can write the above as

$$\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n} = d|\underline{y}\}\} = \sum_{m'} \sum_{\substack{\underline{d}\\(d_{t,n}=d)}} \mathrm{E}_{\underline{n}}\{\beta_t(m)\}\, \mathrm{E}_{\underline{n}}\{\gamma_t(m',m)\}\, \mathrm{E}_{\underline{n}}\{\alpha_{t-1}(m')\}$$

We develop each of these terms independently, so

$$\mathrm{E}_{\underline{n}}\{\beta_t(m)\} = \sum_{m'} \mathrm{E}_{\underline{n}}\{\beta_{t+1}(m')\}\, \mathrm{E}_{\underline{n}}\{\gamma(m,m')\}$$

and

$$\mathrm{E}_{\underline{n}}\{\alpha_{t-1}(m')\} = \sum_{m} \mathrm{E}_{\underline{n}}\{\alpha_{t-1}(m)\}\, \mathrm{E}_{\underline{n}}\{\gamma(m,m')\}$$

As $\gamma(m,m')$ is independent of β_{t+1} and α_{t-1} then the expectations of the individual variables can be determined. The process is therefore recursive as in the MAP decoding algorithm.

We can now calculate the expected value of the transition metric over the input noise distribution

$$\begin{aligned}\mathrm{E}_{\underline{n}}\{\gamma_t(m,m')\} &= \mathrm{E}_{\underline{n}}\{p(y_{t,n=1}|d_{t,n=1} = d_1)p(y_{t,n=2}|d_{t,n=2} = d_2)\cdots p(y_{t,n=r}|d_{t,n=r} = d_r)\} \\ &= \prod_{n=1}^{r} \mathrm{E}_{\underline{n}}\{p(y_{t,n}|d_{t,n} = d_n)\}\end{aligned}$$

where $d_n \in \{-1,+1\}$. The conditional probability for a given coded symbol ($d_{t,n} = -1$, where $y_{t,n} = -1 + n_t$ because the all zero sequence is sent) is

$$\begin{aligned} p(y_{t,n}|d_{t,n} = -1) &= \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{(y_{t,n} - d_{t,n})^2}{2\sigma_i^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{n_t^2}{2\sigma_i^2}\right\} \end{aligned}$$

The expectation of this is

$$\begin{aligned} \mathrm{E}_{\underline{n}}\{p(y_{t,n}|d_{t,n} = -1)\} &= \frac{1}{\sqrt{2\pi}\sigma_i}\mathrm{E}_{\underline{n}}\left\{\exp\left\{-\frac{n_t^2}{2\sigma_i^2}\right\}\right\} \\ &= \frac{1}{2\pi\sigma_i^2}\int_{-\infty}^{\infty}\exp\left\{-\frac{n_t^2}{\sigma_i^2}\right\}dn_t \\ &= \frac{1}{2\sigma_i\sqrt{\pi}} \end{aligned} \tag{6}$$

The conditional probability for a given coded symbol ($d_{t,n} = +1$, where $y_{t,n} = -1 + n_t$) is

$$\begin{aligned} p(y_{t,n}|d_{t,n} = +1) &= \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{(y_{t,n} - d_{t,n})^2}{2\sigma_i^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{4 - 4n_t + n^2}{2\sigma_i^2}\right\} \end{aligned}$$

taking expectations we get

$$\mathrm{E}_{\underline{n}}\{p(y_{t,n}|d_{t,n} = +1)\} = \frac{1}{2\sigma_i\sqrt{\pi}}\exp\left\{-\frac{1}{\sigma_i^2}\right\} \tag{7}$$

This analysis described above and the use of (5) gives us a method to calculate the expectation of the conditional channel output probabilities $\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n} = d|\underline{y}\}\}$.

5 Output Variance

5.1 Numerical Evaluation

Now that we have numerically calculated the expectation of the output probability $\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n} = d|\underline{y}\}\}$ in terms of the input variance (σ_i^2) we desire to define the output variance (σ_o^2) in terms of the expectation of the output probabilities. We need to define the MAP decoder output in terms of its variance if we are to use this analytical information as input to another MAP decoder, in a similar way to turbo decoding. To do this we take the likelihood ratio (LR) of the two possible expectations with the expectation of the all zero path on the numerator. This gives

$$\lambda(t) = \frac{\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n} = -1|\underline{y}\}\}}{\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n} = +1|\underline{y}\}\}}$$

which after interleaving (in a similar way to [1]) is equal to

$$\lambda(t)' \;=\; \frac{\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n}=-1\}\}}{\mathrm{E}_{\underline{n}}\{\Pr\{d_{t,n}=+1\}\}} \tag{8}$$

Assuming the output model as described in Section 2 we can rewrite (8). From (6) and (7) we know what this equals and can denote the LR now as

$$\begin{aligned}\lambda(t)' &= \frac{\mathrm{E}_{\underline{n}}\{\Pr\{y'|d=-1\}\}}{\mathrm{E}_{\underline{n}}\{\Pr\{y'|d=+1\}\}} \\ &= \frac{\frac{1}{2\sigma_o\sqrt{\pi}}}{\frac{1}{2\sigma_o\sqrt{\pi}}\exp\left\{-\frac{1}{\sigma_o^2}\right\}} \\ &= \frac{1}{\exp\left\{-\frac{1}{\sigma_o^2}\right\}}\end{aligned}$$

therefore knowing the likelihood ratio of the expectations we can evaluate the output variance.

5.2 Simulation Evaluation

For the simulation result the output variance is evaluated slightly differently. We use (2) and take the log likelihood ratio (LLR) of the two possible conditional probabilities with the all zero path on the numerator. This gives

$$\Lambda(t) \;=\; \log\left\{\frac{\Pr\{d_{t,n}=-1|\underline{y}\}}{\Pr\{d_{t,n}=+1|\underline{y}\}}\right\}$$

which after interleaving is equal to

$$\Lambda(t)' \;=\; \log\left\{\frac{\Pr\{d_{t,n}=-1\}}{\Pr\{d_{t,n}=+1\}}\right\} \tag{9}$$

Using the output model as described in Section 2 we can rewrite (9) as

$$\begin{aligned}\Lambda(t)' &= \log\left\{\frac{\Pr\{y'|d'=-1\}}{\Pr\{y'|d'=+1\}}\right\} \\ &= \frac{2}{\sigma_o^2}(1-n')\end{aligned}$$

taking expectations of both sides we get

$$\begin{aligned}\mathrm{E}_{n'}\{\Lambda(t)'\} &= \mathrm{E}_{n'}\left\{\frac{2}{\sigma_o^2}(1-n')\right\} \\ &= \frac{2}{\sigma_o^2}\end{aligned}$$

therefore knowing the mean LLR we can determine the variance of the noise in the MAP decoder output signal.

6 Simulation Configuration

For the comparison between the simulation and the numerical analysis result we set our information bit block size to $T = 100$, using a non-systematic rate $R = 1/2$ code with generator polynomials $G[5_8, 7_8]$ with memory 2.

To generate the simulation result we calculate (2) for the entire block, with no knowledge of the start or ending states. This is done to match the performance of the analysis which assumes the same thing. We compute these values for an ensemble set of 100 experiments, we then use the method as described in Section 5.2 to evaluate the output variance.

In the numerical analysis result we determine (5) for $E_{\underline{n}}\{\Pr\{d_{50,1} = d|\underline{y}\}\}$. We then follow the method as described in Section 5.1 to evaluate the output variance.

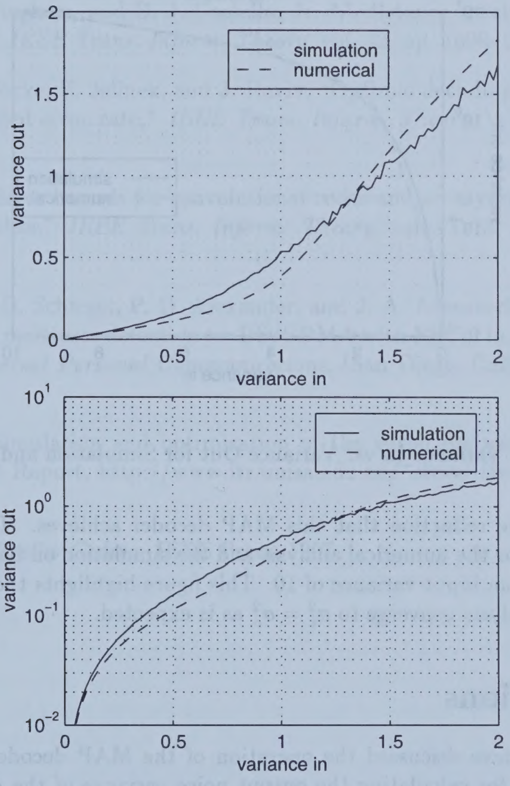


Figure 1: Variance In vs. Variance Out for Simulation and Analysis.

The plots in Fig. 1 shows the results for the numerical analysis and the simulation on both a log scale and a linear scale, up to an input variance of 2. This figure highlights

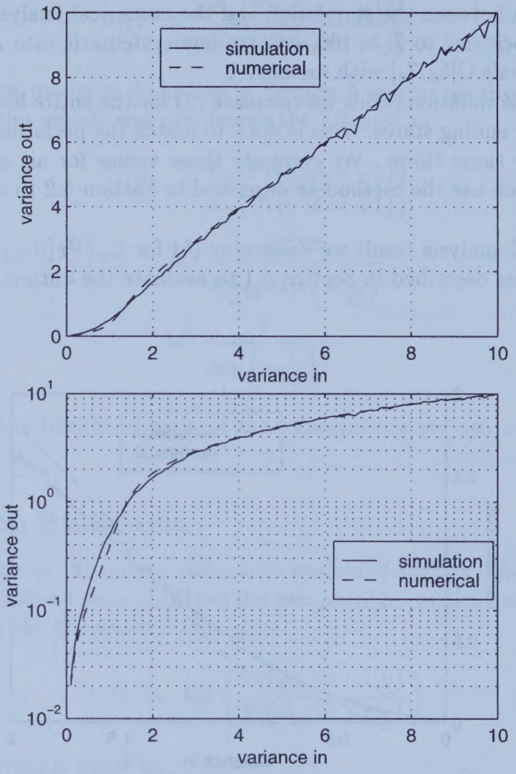


Figure 2: Variance In vs. Variance Out for Simulation and Analysis.

the output variance reduction that this MAP decoder achieves. The plots in Fig. 2 shows the results for the numerical analysis and the simulation on both a log scale and a linear scale, up to an input variance of 10. This figure highlights the fact that both the simulation and analysis converge to $\sigma_i^2 = \sigma_o^2$ as is expected.

7 Conclusions

In this paper we have discussed the operation of the MAP decoder before deriving a numerical method for calculating the output noise variance of the decoder in terms of the input noise variance of the decoder. This result is important as it will allow us to analyse systems that incorporate MAP decoders in an iterative process, which includes iterative multiuser receivers and turbo codes.

The result discussed is a numerical result, a closed form result isn't attempted. The analysis however needs only to know the generator polynomials and the memory of the code to determine its output noise variance as a function of its input noise variance.

References

[1] C. Berrou, A. Glavieux, and P. Thitmajshima, "Near Shannon limit error-correcting coding and decoding:turbo-codes," in *IEEE Int. Conf. on Communications*, (Geneva, Switzerland), pp. 1064–1070, May 1993.

[2] S. Benedetto and G. Montorsi, "Unveiling turbo codes: Some results on parallel concatenated coding schemes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 409–428, March 1996.

[3] S. Benedetto and G. Montorsi, "Performance of turbo codes," *IEE Electron. Lett.*, vol. 31, pp. 163–165, Feb. 1995.

[4] L. C. Perez, J. Seghers, and D. J. Costello, Jr, "A distance spectrum interpretation of turbo codes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1698–1709, Nov. 1996.

[5] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimising symbol error rate," *IEEE Trans. Inform. Theory*, vol. 20, pp. 284–287, 1974.

[6] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.

[7] M. C. Reed, C. B. Schlegel, P. D. Alexander, and J. A. Asenstorfer, "Reduced complexity iterative multiuser detection for DS/CDMA with FEC," in *International Conference on Universal Personal Communications*, (San Diego, California), Oct. 1997. To Appear.

[8] F. Malardel, "Simulation and optimisation of the turbo decoding algorithm." Final Year Project Report, http://www.itr.unisa.edu.au/~steven/turbo/fm.ps.gz, Nov. 1996.

[9] C. B. Schlegel, *Trellis Coding*. IEEE Press, 1997.

CURVES WITH MANY POINTS

by

Ian F. Blake
Hewlett-Packard Laboratories
1501 Page Mill Road, Palo Alto, CA 94304

ABSTRACT

I. INTRODUCTION

The use of curves in algebraic geometries to construct codes and sequences over finite fields has produced remarkable results. Certainly the most notable result in algebraic coding theory in the last two decades is the explicit construction of asymptotically good codes, i.e. those that achieve the rate/minimum distance curve of the Varshamov-Gilbert bound using curves in algebraic geometries. Important aspects of these developments include the determination of the number of rational points on the curve, which determines the length of the code for a given finite field F_q, the dimension of the subspaces defined in terms of bivariate bases and the minimum distance of the code.

A problem of both practical and theoretical interest is the determination of curves containing a number of rational points close to the maximum possible, as determined by the Hasse-Weil bound. There has been considerable attention to this problem in recent years with significant progress achieved. This note briefly surveys some of the directions that have proven most interesting. The work can be viewed as an addendum to the excellent paper of Garcia and Stichtenoth [6] on this subject.

II. PRELIMINARIES.

A very brief introduction to the background needed to address the problem of constructing curves with many points is given. Introductions to the material are given in [1], [6] and [24].

Let F_q be a finite field and $\bar{F}_q$ its algebraic closure. Let $f(x, y) \in F_q[x]$ be a bivariate polynomial over F_q and consider the set

$$\mathcal{C} = \{(u, v) \in \bar{F}_q \times \bar{F}_q | f(u, v) = 0\}$$

i.e. the set of solutions of the equation $f(x, y) = 0$ over $\bar{F}_q$. Each such solution is referred to as a point of the curve. The subset of these points with coordinates in F_q is referred to as the subset of rational points. We will only be interested in the case where $f(x, y)$ is an irreducible polynomial over F_q and where the curve is smooth i.e. contains no non-singular points, points where at least one of the formal derivatives vanishes.

There is an entirely equivalent approach to the subject using function fields [24]. An extension F of F_q is an algebraic function field over F_q if there exists an element $x \in F$ which is transcendental over F_q so that F is a finite extension of $F_q(x)$, the rational field of F_q, and F_q is algebraically closed in F. It is possible then to find an element $y \in F$ such that $F = F_q(x, y)$ which gives an irreducible polynomial $f(x, y) \in F_q[x, y]$. As before, such a polynomial defines a curve. The terminology will be used interchangeably here.

Associated with a function field or curve, is an invariant g, the genus, a positive integer, related to the dimension of certain subspaces associated with divisors of the curve ([24], p. 21). This parameter is often difficult to evaluate for a given curve although a variety of bounds for it are available ([24], pp. 129-133). For example, the case where y satisfies the equation

$$y^n + f_1(x)y^{n-1} + \cdots + f_{n-1}(x)y + f_n(x) = 0\ ,\quad f_j(x) \in F_q[x],\ \deg f_j \leq j\ ,$$

then $g \leq (n-1)(n-2)/2$, and satisfies this with equality iff the corresponding projective plane curve is non-singular. For example, if $F = F_q(x)$, the rational function field of F_q has genus 0.

Determining the number of rational points on a curve of genus g is often a challenging problem. To obtain information on it, define the zeta function of the function field F as

$$Z(t) = \sum_{n \geq 0} A_n t^n$$

where A_n is the number of positive divisors of the field of degree n. It can be shown that this function is of the form

$$Z(t) = \frac{L(t)}{(1-t)(1-qt)}$$

where $L(t)$ is the L polynomial of the curve and

$$\begin{aligned} L(t) &= \sum_{i=0}^{2g} a_i t^i \in \mathbf{Z}[t],\ a_0 = 1, a_{2g} = q^g, a_{2g-i} = q^{g-i} a_i \\ &= \prod_{i=0}^{2g} (1 - \alpha_i t)\ . \end{aligned}$$

It can be shown that $|\alpha_i| = q^{1/2}$, $i = 1, 2, \cdots, 2g$, a very important result referred to as the Hasse-Weil theorem. It readily follows that

$$|N_1 - (q+1)| \leq 2gq^{1/2} \tag{1}$$

which is also referred to as the Hasse-Weil theorem, where N_1 is the number of rational points of the curve. Serre [22] has sharpened this upper bound to

$$|N_1 - (q+1)| \leq g[2q^{1/2}]$$

where $[x]$ is the integer part of x.

If N_r denotes the number of points of the curve with coordinates in F_{q^r} (the curve defined over F_q) then

$$N_r = q^r + 1 - \sum_{i=1}^{2g} \alpha_i^r\ .$$

Thus the number of solutions (of the curve defined over F_q) over F_{q^r} is entirely determined by $N_1, N_2, \cdots, N_g$.

We define a curve of genus g over F_q with N_1 points to be *maximal* if $N_1 = q+1+2gq^{1/2}$. To achieve this upper bound it is clearly necessary that q be a square, $q = r^2$. Let $N_q(g)$ be the maximum number of points on any curve of genus g over F_q, $N_q(g) \leq q + 1 + 2gq^{1/2}$. A curve with $N_q(g) < q + 1 + 2gq^{1/2}$ points will be called *optimal.* A curve that contains "almost" $q + 1 + 2gq^{1/2}$ points will be referred to as "a curve with many points". Subsequent sections survey recent contributions of the past few years to the problem of constructing curves with many points. There is overlap of this work with that of [6], to which this work is a supplement, for the sake of readability.

III. MAXIMAL AND OPTIMAL CURVES.

Let $N_q(g)$ be the largest integer such that there exists an elliptic curve over F_q with $N_q(g)$ points. A curve with $N = N_q(g) = q+1+2q^{1/2}$ points is called *maximal.* A curve with $N = N_q(g) < q+1+2q^{1/2}$ is called *optimal.* The important properties of maximal and optimal curves are given in [6] and are briefly noted here.

Clearly for a curve to be maximal it is necessary that q be a square, $q = r^2$. It is known that if $g > (q - q^{1/2})/2$ the curve cannot be maximal i.e. $N_q(g) < q+1+2gq^{1/2}$. Thus for a maximal curve, $g \leq r(r-1)/2$, $q = r^2$. If a curve has genus $g = r(r-1)/2$, then it is maximal and any such curve is a Hermitian curve [19] and must necessarily have r^3+1 points. Also, any maximal curve over F_q, $q = r^2$ must have genus of the form $g \leq (r-1)^2/4$ or $g = rm/2$, $(r-1)/2 \leq m \leq r-1$ [31].

The L function of a maximal function field or curve, is of the form $(1+q^{1/2}t)^{2g}$ [24]. Since all subfields of a maximal function field are maximal, the L function of subfields are also of this form and therefore divide the parent L function.

Beyond maximal curves, considerable effort has been made to determine the value of $N_q(g)$ when it is less than $q+1+2gq^{1/2}$, for either small values of g or q. These are briefly reviewed [6] and some more recent results stated.

For the case of genus 1 (elliptic curves), the story is quite complete [22] credited to [28], [29]. For $q = p^a$ if a is an odd integer, $a \geq 3$ and $p|[2q^{1/2}]$, then q is called exceptional. Then

$$N_q(1) = \begin{cases} q+1+[2q^{1/2}], & q \text{ nonexceptional} \\ q+[2q^{1/2}], & q \text{ exceptional.} \end{cases}$$

There also exists necessary and sufficient conditions [28] for the existence of a curve with exactly N points when the characteristic does not divide $(N-q-1)$.

The case of $g = 2$ was considered by Serre [22], [23]. If q is a nonsquare, $q = p^a$, a odd, call q special if either $p|[2q^{1/2}]$ or if q is representable over the integers by x^2+1, x^2+x+1 or x^2+x+2. If q is nonsquare and nonspecial then

$$N_q(2) = q+1+2[2q^{1/2}].$$

If q is special then

$$N_q(2) = \begin{cases} q+2[2q^{1/2}], & \text{if } \{2q^{1/2}\} > (\sqrt{5}-1)/2 \\ q-1+2[2q^{1/2}], & \text{if } \{2q^{1/2}\} < (\sqrt{5}-1)/2. \end{cases}$$

where $\{x\}$ is the fractional part of x. Thus $N_q(2)$ is known in all cases.

For the case of $g = 3$, much less is known. Many isolated examples of maximal curves of genus 3 exist. Van der Geer and van der Vlugt [27] give the following infinite class, based on considerations from coding theory: For $q = 2^m$, m even, $m \geq 6$, there exists a maximal curve with genus 3 over F_q, giving a curve with $q+1+6q^{1/2}$. Examples of other infinite classes of maximal curves from that work include: for

$$q = 2^m,\ m \equiv 2 \bmod 4, \quad 1 \leq a \leq m/2$$

there exists a maximal curve of genus $2^a - 1$. Similarly, for

$$q = 2^m,\ m \text{ even},\ m \geq 4,\ 1 \leq a \leq m/2,$$

there exists a maximal curve with genus $(2^r - 1)q^{1/2}/2$. Other examples of non-Hermitian maximal and optimal function fields are given in [10], [27] and elsewhere in the literature. Table 1 attempts to summarize known results on optimal curves, gleaned from the literature. Many of the entries have been determined by explicitly constructing a curve with the shown number of points and then proving no curve with more points exist. The lower bounds are likewise determined by explicit construction of a curve. The entries with ranges arose mainly by using an upper bound technique of Serre [23], [16], [12].

IV. ASYMPTOTICS.

Define the quantity

$$A(q) = \limsup_{g \to \infty} \frac{N_q(g)}{g}$$

and notice from the Hasse-Weil bound that $A(q) \leq 2q^{1/2}$ or, using the Serre improvement, $A(q) \leq [2q^{1/2}]$. In connection with coding theory, any sequence of curves C_i, of genus g_i, respectively, for which

$$\limsup_{g \to \infty} \frac{N_i}{g_i} > 0$$

is called asymptotically good, and asymptotically bad otherwise. Ihara [9] showed that $A(q) \leq (\sqrt{8q+1} - 1)/2$ and Ihara and Tsfasman et al [26] showed that when q is a square, $A(q) = q^{1/2} - 1$. Drinfeld and Vladut [2] showed that for any q, $A(q) \leq q^{1/2} - 1$.

Serre [23] showed that $A(q) \geq c \log(q)$. Zink [33] (see also [21]) showed that for p a prime,

$$A(p^3) \geq \frac{2(p^2-1)}{p+2}$$

by exhibiting a suitable infinite family of modular curves. Perret [18] showed that if l is a prime number and k a positive integer, then if $q = p^e$ is a primitive kth root of unity in F_l then

$$\begin{aligned} A(q^l) &\geq \frac{\sqrt{l(q-1) - 2l}}{l-1} \quad \text{if } k = 1 \\ A(q^k) &\geq \frac{\sqrt{l(q-1) - 2l}}{l-1} \quad \text{if } k \geq 2\,. \end{aligned}$$

With this approach he also defined a family of asymptotically good curves. though not as good as those of Tsfasman et al [26]. Schoof [21] showed that $A(2) \geq 2/9 \cong .222$ a bound apparently found earlier by Serre. He established this result by constructing a field tower, a notion discussed in the next section. Stichtenoth [24] combines these results to show that

$$\begin{aligned} .222 &\leq A(2) \leq .414 \\ .333 &\leq A(3) \leq .732 \\ .500 &\leq A(5) \leq 1.236 \end{aligned}$$

It is noted [6] that for any value of q for which q is not a square, the exact value of $A(q)$ is not known.

The next section briefly describes two asymptotically good towers of function fields.

V. TOWERS OF FUNCTION FIELDS.

As noted, it remains a difficult problem to construct an asymptotically good sequence of function fields or curves. Two such remarkable sequences have recently been constructed and their properties are briefly outlined here. In both cases an infinite sequence of function fields $F_1 \subseteq F_2 \subseteq \cdots$ over F_{q^2} is constructed such that

$$\lim_{i \to \infty} \frac{N(F_i)}{g(F_i)} = q - 1$$

i.e. the Drinfeld-Vladut bound is achieved.

For the first tower [4] set $F_1 = F_{q^2}(x_1)$, the rational function field over F_{q^2}. For $n \geq 1$ set

$$F_{n+1} = F_n(z_{n+1})$$

where

$$z_{n+1}^q + z_{n+1} = x_n^{q+1} \quad , \quad x_n = z_n/x_{n-1} \in F_n \ , \ n \geq 2.$$

Notice that F_2 is simply the Hermitian function field and that $[F_{n+1} : F_n] = q,\ F_n|F_1 = q^{n-1}$. The solutions to these [4] equations are studied and it is shown that the genus of F_n over F_1 is given by

$$g(F_n) = \begin{cases} q^n + q^{n-1} - q^{(n+1)/2} - 2q^{(n-1)/2} + 1 & if\ n \equiv 1 \bmod 2 \\ q^n + q^{n-1} - \frac{1}{2}q^{n/2+1} - q^{n/2-1} + 1 & if\ n \equiv 0 \bmod 2 \end{cases}$$

The number of points on F_n over F_{q^2} of degree one (rational points) for $n \geq 3$ is

$$N(F_n) \geq (q^2 - 1)q^{n-1} + q$$

with equality if $n \geq 3$ and $q \equiv 1 \bmod 2$. Taking the first two terms of $g(F_n)$ it is clear that asymptotically

$$\frac{N(F_n)}{g(F_n)} \sim \frac{(q+1)(q-1)q^{n-1}}{q^{n-1}(q+1)} \to q - 1$$

The second tower [5], [17] is constructed in a similar manner, again over F_{q^2}, with slightly different equations:

$$z_{n+1}^q + z_{n+1} = \frac{z_n^q}{z_n^{q-1} + 1} \quad , n \geq 1.$$

It is shown that the genus of F_n is given by the expressions:

$$g(F_n) = \begin{cases} (q^{n/2} - 1)^2 & if\ n \equiv 0 \bmod 2 \\ (q^{(n+1)/2} - 1)(q^{(n-1)/2} - 1) & if\ n \equiv 1 \bmod 2 \end{cases}$$

and that the number of rational points of F_n is at least $(q^2 - q)q^{n-1}$. It follows readily that $N(F_n)/g(F_n) \to q - 1$. The structure and properties of this tower is investigated in some detail in [6], [17].

VI. OTHER CLASSES OF CURVES.

There is an extensive literature on other classes of curves, generally neither optimal nor maximal. Two such classes are briefly mentioned here.

Wolfmann [30] considered the class:

$$y^q - y = ax^s + b \ , \ \ a \in F_{q^k}^*,\ b \in F_{q^k},\ (s, q) = 1.$$

It can be shown that the genus of all such curves is $g = (q-1)(s-1)/2$. Assume that $k = 2t$, $s|(q^k - 1)$, $ns = q^k - 1$. If there exists a divisor r of t such that $q^r \equiv -1 \bmod s$, then let $\epsilon = (-1^{t/r}$, $\epsilon_1 = \epsilon^u$, $us = q^r + 1$ then the number of solutions is given by the following:

$$\begin{array}{ll}
\text{if } a^n = \epsilon_1 \text{ and } tr(b) = 0 \text{ then :} & N = q^{2t} + 1 - \epsilon(q-1)(s-1)q^t \\
\text{if } a^n = \epsilon_1 \text{ and } tr(b) \neq 0 \text{ then :} & N = q^{2t} + 1 + \epsilon(s-1)q^t \\
\text{if } a^n \neq \epsilon_1 \text{ and } tr(b) = 0 \text{ then :} & N = q^{2t} + 1 - \epsilon(q-1)q^t \\
\text{if } a^n \neq \epsilon_1 \text{ and } tr(b) \neq 0 \text{ then :} & N = q^{2t} + 1 - \epsilon q^t
\end{array}$$

further conditions are given under which the curves are maximal [30]. The work illustrates the ingenuity of many of the constructions and analyses.

Garcia [8] considers the equation

$$y^n = f(x)$$

over F_q, $\text{char}(q) = p > 2$. It is shown that if $f(x)$ is separable and $r = \deg f$ and n are relatively prime, then the genus of the curve is $g = (n-1)(r-1)/2$ and the number of points on the curve is upper bounded by

$$N \leq \begin{cases} (n-1)(r-1)/2 - 1 + r(q+2)/2 & r > n \\ (n-1)(r-1)/2 - 1 + n(q+2)/2 & r < n. \end{cases}$$

The first bound is tighter than the Hasse-Weil bound if $q^{1/2}/2 < n < r < q^{1/2}$ and the second bound is tighter if $q^{1/2}/2 < r < n < q^{1/2}$.

Numerous other interesting classes of curves can be found in the literature. The papers of Niederreiter and Xing are particularly rich in such examples.

VII. COMMENTS.

Recent years have seen interesting avenues of development in the construction of curves with many points. The determination of asymptotically good curves over lower field sizes is an especially interesting pursuit as is the search for maximal, non-Hermitian curves. This work has attempted to update the more general treatment of curves in [6].

References

[1] R.E. Blahut, *Lectures in Algebraic Coding Theory*, preprint, July, 1997.

[2] V.G. Drinfeld and S. Vladut, Sur le nombre de points d'une courbe algébrique, *Anal. Fonct. et Appl.*, vol. 17, pp. 68-69, 1983.

[3] W. Fulton, *Algebraic Curves: An Introduction to Algebraic Geometry*, Reading, MA: Addison-Wesley Publishing Company, 1969.

[4] A. Garcia and H. Stichtenoth, A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound, *Inventiones Math.*, vol. 121, pp. 211-222, 1995.

[5] A. Garcia and H. Stichtenoth, On the asymptotic behaviour of some towers of function fields over finite fields, *Jour. Number Theory*, vol. 61, pp. 248-273, 1996.

[6] A. Garcia and H. Stichtenoth, Algebraic function fields over finite fields with many rational points, *IEEE Trans. Information Theory*, vol. 41, pp. 1548-1563, 1995.

[7] A. Garcia and J.F. Voloch, Fermat curves over finite fields, *Jour. Number Theory*, vol. 30, pp. 345- 356, 1988.

[8] A. Garcia, The curves $y^n = f(x)$ over finite fields, *Arch. Math.*, vol. 54, pp. 36-44, 1990.

[9] Y. Ihara, Some remarks on the number of rational points of algebraic curves over finite fields, *J. Fac. Sci. Tokyo*, vol. 28, pp. 721-724, 1981.

[10] G. Lachaud, Sommes d'Eisenstein et nombre de points de certaines courbes algébriques sur les corps finis, *C.R. Acad. Sci., Paris, Série I*, vol. 305, pp. 729-732, 1987.

[11] Yu. I. Manin, What is the maximum number of points on a curve over F_2, *J. Fac. Sci. Univ. Tokyo, Sec. IA*, vol. 28, pp. 715-720, 1981.

[12] O. Moreno, D. Zinoviev and V. Zinoviev, On several new projective curves over F_2 of genus 3, 4 and 5, *IEEE Trans. Information Theory*, vol. 41, pp. 1643-1647, 1995.

[13] H. Niederreiter and C. Xing, Low-discrepancy sequences and global function fields with many rational places, *Finite Fields and their Applications*, vol. 2, pp. 241-273, 1996.

[14] H. Niederreiter and C. Xing, Cyclotomic function fields, Hilbert class fields, and global function fields with many rational places, *Acta Math.*, vol. 79, pp. 53-76, 1997.

[15] H. Niederreiter and C. Xing, Explicit global function fields over the binary field with many rational places, *Acta Math.*, vol. 75, pp. 383-396, 1996.

[16] H. Niederreiter and C. Xing, Drinfeld modules of rank 1 and algebraic curves with many rational points II, *Acta Math.*, vol. 81, pp. 81-100, 1997.

[17] R. Pelikaan, H. Stichtenoth and F. Torres, Weierstrass semigroups in an asymptotically good tower of function fields, preprint.

[18] M. Perret, Sur le nombre de points d'une courbe sur un corps fini; application aux codes correcteurs d'erreurs, *C.R. Acad. Sci., Paris, Série I*, vol. 309, pp. 177-182, 1989.

[19] H.-G. Rück and H. Stichtenoth, A characterization of Hermitian function fields over finite fields, *J. reine angew. Math.*, vol. 457, pp. 185-188, 1994.

[20] R. Schoof, Families of curves and weight distributions of codes, *Bull. Amer. Math. Soc.*, vol. 32, pp. 171-183, 1996.

[21] R. Schoof, Algebraic curves over F_2 with many rational points, *Jour. Number Theory*, vol. 1, pp. 6-14, 1992.

[22] J.-P. Serre, Nombres des points des courbes algébriques sur F_q, *Séminaire de Théorie des Nombres de Bordeaux*, 1982/83, no. 22.

[23] J.-P. Serre, Sur le nombre des points rationnels d'une courbe algébrique sur un corps fini, *C.R. Acad. Sci., Paris, Série I*, vol. 296, pp. 397-402, 1983.

[24] H. Stichtenoth, *Algebraic Function Fields and Codes*, Berlin: Springer-Verlag, 1991.

[25] K.-O Stöhr and J.F. Voloch, Weierstrass points and curves over finite fields, *Proc. London Math. Soc.*, vol. 52, pp. 1-19, 1986.

[26] M.A. Tsfasman, S.G. Vladut and T. Zink, Modular curves, Shimura curves and Goppa codes better than the Varshamov-Gilbert bound, *Math. Nach.*, vol. 109, pp. 21-28, 1982.

[27] G. van der Geer and M. van der Vlugt, Curves over finite fields of characteristic 2 with many rational points, *C.R. Acad. Sci., Paris, Série I*, vol. 317, pp. 593-507, 1993.

[28] W.C. Waterhouse, Abelian varieties over finite fields, *Ann. Sci. E.N.S.*, vol. 2, pp. 521-560, 1969.

[29] A. Weil, *Variétés Abéliennes et courbes algébriques*, Paris: Hermann, 1948.

[30] J. Wolfmann, The number of points on certain algebraic curves over finite fields, *Communications in Algebra*, vol. 17, pp. 2055-2060, 1989.

[31] C. Xing and H. Stichtenoth, The genus of maximal function fields, *Manuscripta Math.*, vol. 86, pp. 217-224, 1995.

[32] C. Xing and H. Niederreiter, Modules de Drinfeld et courbes algébriques ayant beaucoup de points rationnels, *C.R. 156 Acad. Sci., Paris, Série I*, vol. 322, pp. 651- 654, 1996.

[33] Th. Zink, Degeneration of Shimura surfaces and a problem in coding theory, in *Lecture Notes in Computer Science*, vol. 199, L. Budach ed., Berlin: Springer-Verlag, 1986.

$g \setminus q$	2	3	4	5	7	8	9	11	13	16	17	19
1	5	7	9	10	13	14	16	18	21	25	26	28
2	6	8	10	12	16	18	20	24	26	33	32	36
3	7	10	14	16	20	24	28	28	32	38	40	44
4	8	12	15	18		25-29				45-49		
5	9	12*	17-18	20*		26-34				45-57		
6	10	14*	20	21*		33*						
7	10	16*	21-22	20*		33*						
8	11	15*	21-24	22*								
9	12	19	26	24*								
10	13	19*	27-28	26*								
11	13-14	20*	25-30	26*								
12	14-15	22*	28-31	30*								
13	15	24*	33									
14	15-16	24*	33-35									
15	17	28*	33-37									
16	16-18		36-38									
17	17-18		40									
18	18-19		34-42									
19	20		36-43									
20	19-21		36-45									
21	21		41-47									
22			40-48									
23			40-50									
24			42-52									
25			51-53									

Table 1. Values and ranges for $N_q(g)$.
(From [6], [8], [13], [14], [15], [16], [22], [23], [12], [31])
(Note: A * indicates a lower bound only.)

The Applications of Generalized Bezout's Theorem to the Codes from the Curves in High Dimensional Spaces

Gui-Liang Feng, Junmei Zhu, Xiaofa Shi and T. R. N. Rao *

November 21, 1997

Abstract

In this paper, we generalize Bezout theorem to p polynomials of n variables. This generalized Bezout theorem in higher dimension provides us a new method of estimating the upper bound of the number of intersection points of two or more polynomials curves. Using this bound, the lower bound on the minimum distances of linear codes defined on the curves in high dimensional spaces are obtained. As an application of this theorem, more efficient linear codes defined on the curves in high dimensional spaces are constructed, which are better than the AG codes and the improved AG codes on the same curves.

Index Terms: Bezout's theorem, minimum distance, generalized Hamming weights, algebraic-geometric codes, linear codes.

1 Introduction

For error-correcting codes, the minimum distance is one of most important parameters. It is used to measure the code's capacity of correcting errors or detecting errors or both [3]. The minimum distance d of a linear code C is defined by

$$d = \min_{\substack{\mathbf{u},\mathbf{v}\in C \\ \mathbf{u}\neq\mathbf{v}}} \{d(\mathbf{u},\mathbf{v})\},$$

where $d(\mathbf{u},\mathbf{v})$ expresses the Hamming distance between $\mathbf{u}$ and $\mathbf{v}$.

For an $[n,\ k]$ linear code, we can consider its generalized Hamming weights, which are the generalization of minimum distance and defined as follows:

Definition 1.1 *For any code C of block length n over $GF(q)$, define the* support $\mathcal{X}(C)$ *by*

$$\mathcal{X}(C) = \{i \mid c_i \neq 0 \ \text{ for some } (c_1, c_2, \cdots, c_n) \in C\},$$

---

*This work was supported in part by the National Science Foundation under Grant NCR-9505619.

and the support weight $w_s(C)$ *by*

$$w_s(C) = |\mathcal{X}(C)|.$$

Let C *is a linear* $[n,\ k]$ *code over* $GF(q)$. *For any* r *with* $1 \leq r \leq k$, *the* r-th generalized Hamming weight *of* C *is defined as*

$$d_r(C) = \min\{w_s(D) \mid D \text{ is a } r\text{-dimensional subcode of } C\}.$$

The weight hierarchy of code C *is defined as the set of generalized Hamming weights* $\{d_1(C), d_2(C), \cdots, d_k(C)\}$.

It is easy to see that $d_1(C)$ is the minimum distance or the minimum Hamming weight of code C.

Let C be a q-ary $[n,\ k]$ linear code, we have the following properties of the generalized Hamming weights:

1) (Monotonicity) $1 \leq d_1(C) < d_2(C) < \cdots < d_k(C) \leq n$.

2) (The generalized Singleton bound) $d_H(C) \leq n - k + h$, for $h = 1, 2, \cdots, k$.

These properties were proved for cases $q = 2$ in [4]. When q is a power of any prime, the proof is the same.

Both the determination of the minimum distances and the determination of weight hierarchy for linear codes in full are difficult. A more modest goal is to find acceptable bounds on these weights. The weights of geometric Goppa codes were discussed in [12] and [6]. The bounds on the minimum distance and the generalized Hamming weights of the codes defined on the curves in two-dimensional space were given in [1].

Definition 1.2 *Let* $X = (x_1, x_2, \cdots, x_n)$, $D_{\{f_1, f_2, \cdots, f_p\}}$ *denotes the number of distinct points of the intersection of polynomials* $f_\mu(X) = 0$, *for* $\mu = 1, 2, ..., p$.

Definition 1.3 *Given a sequence of polynomials* $\{f_\mu(X) | \mu = 1, 2, ..., r\}$.

$$D_p^{(r)} = \max\{D_{\{f^*_{\lambda_1}, f^*_{\lambda_2}, \cdots, f^*_{\lambda_p}\}} | \lambda_1, \cdots, \lambda_p \leq r\},$$

where $X = (x_1, x_2, \cdots, x_n)$ *and* $f^*_{\lambda_\mu}$ *expresses a linear combination of* f_i *for i =1, 2, ...,* λ_μ, *and the coefficient of* f_{λ_μ} *is 1, i.e.,* $f^*_{\lambda_\mu} = f_{\lambda_\mu} + \sum_{i=1}^{\lambda_\mu - 1} c_i f_i$.

Let $LS = \{P_1, P_2, \cdots, P_N\}$ be a set of points in the n-dimensional vector space over $GF(q)$, in general, we consider the points in a algebraic curves, i.e., the rational points of a algebraic curve. Let $H_r = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_r\}$ be a set of monomials or polynomials with variables $x_1, x_2, \cdots, x_n$. For every $\mathbf{h}_i$, we have a evaluated vector $(\mathbf{h}_i(P_1), \mathbf{h}_i(P_2), \cdots, \mathbf{h}_i(P_N))^T$. When there is no possibility of any confusion, we also use $\mathbf{h}_r$ to denote its evaluated vector, i.e., we let $\mathbf{h}_i = (\mathbf{h}_i(P_1), \mathbf{h}_i(P_2), \cdots, \mathbf{h}_i(P_N))^T$. Let C_r be the $[N,\ N-r]$ linear code defined by a parity check matrix $\mathbf{H}_r = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_r]^T$. The relationship between $D_p^{(r)}$ and the generalized Hamming weights of linear code C_r is given in the following theorem.

Theorem 1.1 *[1] For a linear code* C_r *defined by* $\mathbf{H}_r$, *i.e., the parity check matrix has* r *rows, if* $D^{(r)}_{r-d^*+h+1} < d^* - 1$, *then the h-th generalized Hamming weight* $d_h(C_r)$ *is at least* d^*, *i.e.,* $d_h(C_r) \geq d^*$.

From the above theorem, we can see that an upper bound on the number of intersection points of two or more polynomials, $D_p^{(r)}$, can be used to obtain the lower bound on the generalized Hamming weight of a linear code.

Bezout's theorem is a useful tool for estimating the number of common roots of two bivariate polynomials. In [1], we generalized Bezout's theorem to several bivariate polynomials. In this paper, we further generalize this result to p polynomials with n variables. This provides us a new approach to estimate the upper bound of the number of intersection points of two or more polynomial curves in high dimensional space. The theorem and its proof is given in the next section. In section 3, we show that, when the polynomials satisfy certain constrains, an upper bound of the number of their common roots can be easily obtained by only considering the leading terms. In section 4, we use the generalized Bezout's theorem to construct some more efficient linear codes. Conclusions are given in section 5.

2 Generalized Bezout's theorem in high dimensional space

Like in the case of p bivariate polynomials, the generalized Bezout's theorem is based on the concept of resultant matrix. Recall that the $x-$*resultant* matrix of p polynomials of x and y is defined as

Definition 2.1 *Let $f_\mu(x,y)$, for $\mu = 1, 2, \cdots, p$, be polynomials in x and y. Without loss of the generality, $deg_x f_1 \geq deg_x f_2 \geq \cdots \geq deg_x f_p$, and let $deg_x f_1 = m$ and $deg_x f_2 = n$, where $deg_x f_\mu$ indicates the maximal i such that the monomial $x^i y^j$ is a term in f_μ. We define the x-resultant matrix of these p curves or polynomials as the following $\Sigma \times (m+n)$ matrix, where $\Sigma = \sum_{\mu=1}^{p} (m + n - deg_x f_\mu)$ and $s = deg_x f_p$:*

$$\begin{bmatrix}
a_0^{(1)} & a_1^{(1)} & . & . & . & a_m^{(1)} & 0 & . & . & 0 \\
0 & a_0^{(1)} & a_1^{(1)} & . & . & . & a_m^{(1)} & 0 & . & 0 \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
0 & 0 & . & 0 & a_0^{(1)} & a_1^{(1)} & . & . & . & a_m^{(1)} \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
a_0^{(p)} & a_1^{(p)} & . & . & a_s^{(p)} & 0 & 0 & . & . & 0 \\
0 & a_0^{(p)} & a_1^{(p)} & . & . & a_s^{(p)} & 0 & . & . & 0 \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
0 & 0 & . & . & . & 0 & a_0^{(p)} & a_1^{(p)} & . & a_s^{(p)}
\end{bmatrix}$$

Let $R(y) = Res_x(f_1, f_2, \cdots, f_p)$ be the non-zero determinant of nonsingular submatrix with the smallest degree of y of the x-resultant matrix. $R(y)$ is called the x-resultant of polynomials $f_1, f_2, \cdots$, and f_p.

Our generalized Bezout's theorem to the case of p polynomials of n variables gives an upper bound of the number of common roots of a set of polynomial equations

$$\begin{cases} f_1(x_1, x_2, \cdots, x_n) = 0 \\ f_2(x_1, x_2, \cdots, x_n) = 0 \\ \cdots\ \cdots\ \cdots\cdots\ \cdots \\ f_p(x_1, x_2, \cdots, x_n) = 0 \end{cases} \tag{1}$$

The systematic approach to obtain this bound, as stated in the theorem, is to eliminate one variable one by one through their corresponding resultant matrix. Assume $R_1(f_1, f_2, \cdots, f_p)$ is the resultant matrix with respect to x_n. By setting the determinate of all the non-singular full-submatrices to zero, we can have another set of polynomial equations, with $n-1$ variables $x_1, x_2, \cdots, x_{n-1}$.

$$\begin{cases} f_1^{(1)}(x_1, x_2, \cdots, x_{n-1}) = 0 \\ f_2^{(1)}(x_1, x_2, \cdots, x_{n-1}) = 0 \\ \vdots \\ f_{p_1}^{(1)}(x_1, x_2, \cdots, x_{n-1}) = 0 \end{cases} \tag{2}$$

Repeat this operation and after $n-2$ iterations, we get

$$\begin{cases} f_1^{(n-2)}(x_1, x_2) = 0 \\ f_2^{(n-2)}(x_1, x_2) = 0 \\ \vdots \\ f_{p_{n-2}}^{(n-2)}(x_1, x_2) = 0 \end{cases} \tag{3}$$

By eliminating x_2 using $x_2 -$ *resultant*, we finally get

$$\begin{cases} f_1^{(n-1)}(x_1) = 0 \\ f_2^{(n-1)}(x_1) = 0 \\ \vdots \\ f_{p_{n-1}}^{(n-1)}(x_1) = 0 \end{cases} \tag{4}$$

Let g be the minimal degree among all these $f_\mu^{(n-1)}(x_1)$. We have the generalized Bezout theorem for p polynomials with n variables.

Theorem 2.1 *The number of common roots of p polynomials equations* $f_\mu(x_1, x_2, \cdots, x_n) = 0$, *for* $\mu = 1, 2, ..., p$, *is at most equal to* g. *The common multiple roots are counted as such.*

The proof of this theorem needs some results from the generalized Bezout's theorem of several bivariate polynomials. Here we give these results without proving them. Detailed proof can be found in [1].

Theorem 2.2 *If* $R(y) = (y - \beta)^r g(y)$ *where* $g(\beta) \neq 0$ *is the determinant of a non-singular full-submatrix of* $x-$*resultant matrix of* $f_\mu(x, y)$, $1 \leq \mu \leq p$, *then* $f_\mu(x, \beta)$ *has at most* r *common roots.*

Theorem 2.3 *The number of common roots of p polynomial equations* $f_\mu(x,y) = 0$*, for* $\mu = 1, 2, ..., p$*, is at most equal to* $degR^*(y)$*, where* $R^*(y)$ *is the* $R(y)$ *whose degree is the smallest. Multiple common roots are counted as such.*

Proof of Theorem 2.1: Using mathematical induction with respect to the number of variables, n.

1. $n = 2$ is just theorem 2.3.

2. Assume that the theorem is true when there are $n-1$ variables, we need to prove it for n variables, that is, the number of solutions for equation 1 is at most g.

By setting the determinant of the non-singular full-submatrices of the x_n-resultant matrix of $f_\mu(x_1, x_2, \cdots, x_n)$ to zero, we get another set of equations, as written in equation 2. Since the equations in 2 have only $n-1$ variables $x_1, \cdots, x_{n-1}$, according to the assumption, they have at most g common roots. Note that we count common multiple order roots as the number of their multiplicity.

Assume that $(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{n-1})$ is a common root of equation 2, and its multiplicity is r. Therefore, the summation of r for all $(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{n-1})$ is at most g. That is,

$$\sum_{(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{n-1})} r = g$$

Since $(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{n-1})$ is a root with multiplicity r, $f_\mu^{(1)}(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{n-2}, x_{n-1}) = 0$, $1 \leq \mu \leq p_1$ must have common factor $(x_{n-1} - \hat{x}_{n-1})^r$. Thus $f_\mu(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{n-2}, \hat{x}_{n-1}, x_n) = 0$, $1 \leq \mu \leq p$ has at most r common roots, as can be seen from theorem 2.2. The total number of roots is the the summation of r for all $(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{n-1})$, which is at most g. The proof is completed. □

3 The application of generalized Bezout's theorem to some restricted equation set: a simple result

As can be seen from last section, given a set of polynomial equations, we can use generalized Bezout's theorem to get an upper bound of their number of common roots. However, the computation can be very intensive when the system is complex. In this section, we show that sometimes, when the polynomials satisfy certain constraints, we can get to the result without much computation, and do not need to follow the procedure step by step.

We first define an order for all monomials, because the final estimation depends only on the leading term of each polynomial equation, as can be seen soon. Lexicographic order is used. In n-dimensional space, where the variables are $x_1, \cdots, x_n$, the order is

$$1 < x_n < x_n^2 < \cdots < x_{n-1} < x_{n-1}x_n < \cdots$$

Each polynomial equation can be represented by a point in the n-dimensional space, if we map it's leading term $x_1^{i_1} \cdots x_n^{i_n}$to the point $(i_1 \cdots i_n)$. The condition that is required to get this simple result is that we are able to pick up n polynomial equations from the set, such that in each axis, there is exactly one point. That is, the j-th polynomial has the leading term of $x_j^{i_j}$ and it contains only variables $x_j, x_{j-1}, \cdots, x_1$. The number of common roots of these n polynomials is at most $i_1^* i_2^* \cdots i_n^*$, the volume of the hypercube defined by these n points.

For the rest polynomial equations, we only consider those whose corresponding point is inside the hypercube, that is, the polynomial having leading term $x_1^{i_1} \cdots x_n^{i_n}$ with $i_j \leq i_j^*$, because those outside points will have no contribution in reducing the number of common roots. An upper bound of the common roots of the first n restricted polynomials and the polynomial with leading term $x_1^{i_1} \cdots x_n^{i_n}$ is

$$i_1^* i_2^* \cdots i_n^* - (i_1^* - i_1)(i_2^* - i_2) \cdots (i_n^* - i_n)$$

This number is just the cardinality of the determine set of $x_1^{i_1} \cdots x_n^{i_n}$ in the n-dimensional space. The determine set of a point $x_1^{i_1} \cdots x_n^{i_n}$ is defined as the integer points within hypercube $[(0, \cdots, 0), (i_1^*, \cdots, i_n^*)]$ but not in $[(i_1, \cdots, i_n), (i_1^*, \cdots, i_n^*)]$.

For more equations with different leading term, an upper bound is the cardinality of the intersection of their corresponding determine set.

We illustrate this result in two-dimensional and three dimensional spaces, when the result is easy to visualize. In two-dimensional space, we have the theorem.

Theorem 3.1 *The number of common roots of equations*

$$\begin{aligned}
g_1(x,y) &= [y^{j_1}]x^{i_1} + f_{12}(y)x^{i_1-1} + \cdots = 0 \\
g_2(x,y) &= [y^{j_2}]x^{i_2} + f_{22}(y)x^{i_2-1} + \cdots = 0 \\
&\vdots \\
g_n(x,y) &= [y^{j_n}]x^{i_n} + \cdots = 0
\end{aligned}$$

where $i_1 > i_2 > \cdots > i_n = 0$ *and* $0 = j_1 < j_2 < \cdots < j_n$, *is at most equal to*

$$j_1 + j_2(i_1 - i_2) + j_3(i_2 - i_3) + \cdots + j_n i_{n-1}$$

The notation $[y^{j_k}]$ *stands for any polynomial with leading term* y^{j_k}.

Proof: Since $i_1 > i_2 > \cdots > i_n$, the x-resultant matrix of the n equations is

$$\begin{bmatrix} RM^{i_2}(g_1) \\ RM^{i_1}(g_2) \\ \vdots \\ RM^{i_1+i_2}(g_n) \end{bmatrix} = \begin{bmatrix}
[y^{j_1}] & f_{12}(y) & \cdot & \cdot & \cdot & * & 0 & \cdot & \cdot & 0 \\
0 & [y^{j_1}] & f_{12}(y) & \cdot & \cdot & \cdot & * & 0 & \cdot & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & \cdot & 0 & [y^{j_1}] & f_{12}(y) & \cdot & \cdot & \cdot & * \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
[y^{j_n}] & 0 & \cdot & \cdot & 0 & 0 & 0 & \cdot & \cdot & 0 \\
0 & [y^{j_n}] & 0 & \cdot & \cdot & 0 & 0 & \cdot & \cdot & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & \cdot & \cdot & \cdot & 0 & 0 & 0 & \cdot & [y^{j_n}]
\end{bmatrix} \tag{5}$$

If we can find an upper triangular submatrix, it is very easy to calculate its determinant. One such matrix can be obtained by choosing all the i_2 rows from $RM^{i_2}(g_1)$, the last (i_1-i_2) rows from $RM^{i_1}(g_2)$, and so on, and finally the last i_{n-1} rows from $RM^{i_1+i_2}(g_n)$. Since the determinant of this submatrix is a polynomial of y with maximum degree of

$$j_1 + j_2(i_1 - i_2) + j_3(i_2 - i_3) + \cdots + j_n i_{n-1}$$

this number is an upper bound on the number of common roots. □

Note that the conditions $j_1 = 0$ and $i_n = 0$ guarantee the existence of two points $(i_1, 0)$ and $(0, j_n)$, one in each axis. The order of these polynomials is chosen as such because other polynomials out of this order have no contribution in lowering this bound.

In three dimensional space, we have

Theorem 3.2 *The number of common roots of the following set of equations*

$$x^{i_0} + f_{12}(y,z)x^{i_0-1} + \cdots = 0 \tag{6}$$

$$(y^j z^k + \cdots)x^i + f_{22}(y,z)x^{i-1} + \cdots = 0 \tag{7}$$

$$y^{j_0} + f_{32}(z)y^{j_0-1} + \cdots = 0 \tag{8}$$

$$z^{k_0} + f_{42}z^{k_0-1} + \cdots = 0 \tag{9}$$

is $i_0 j_0 k_0 - (i_0 - i)(j_0 - j)(k_0 - k)$.

This theorem can be easily proved by following the procedure of generalized Bezout's theorem: eliminate variables x and y by their corresponding resultant matrices, and find out the order of z.

This is a bound that can be easily calculated, which is simply the cardinality of the determine set, even though it may not be a very tight bound. The determine set of point (i, j, k) in three-dimensional space can be illustrated by the following figure. The number of common roots of more polynomials is upper bounded by the cardinality of the intersection of their corresponding determine sets.

4 Construction of Some Codes in High Dimensional Spaces

In the is section, we are going to give some examples to show how to use Bezout's Theorem and the results in the above sections to construct more efficient codes in high dimensional spaces.

4.1 Linear Codes Better Than the Improved Hermitian Codes

Now we consider irreducible Hermitian space curve over $GF(2^2)$:

$$\begin{cases} x^3 + y^2 + y = 0, \\ y^3 + z^2 + z = 0. \end{cases}$$

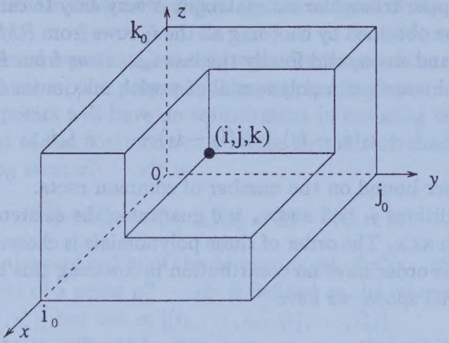


Figure 1: The determine set of point (i,j,k)

From the definition, we have $w(x) = 4, w(y) = 6$ and $w(z) = 9$. Based on the weights, we have the following increasing series of monomials:

$$\{1, x, y, x^2, z, xy, x^3, xz, x^2y, yz, x^2z, x^3y, xyz, x^3z, \cdots\}.$$

If we take seven polynomials as $\{1, x, y, x^2, z, xy, xz + yz\}$ and consider $D_4^{(7)}$, we have the following theorem. The proof is omitted here because of the limited space.

Theorem 4.1 $D_4^{(7)} \leq 3$.

From Riemann-Roch Theorem, the AG code with $d \geq 5$ should have $r = d + g - 1 = 10$. This means that the AG code with $d \geq 5$ is (16, 6, $\geq$ 5). In [10], an improved AG code (16, 8, $\geq$ 5) was given. Using Theorem 4.1, a better AG code can be constructed by taking $\mathrm{H} = [1, x, y, x^2, z, xy, xz + yx]^T$. The new code is linear code (16, 9, $\geq$ 5). It has more information bits and therefore is more efficient.

4.2 More Efficient Double-byte Error-correcting Codes

Consider finite field $GF(q^3) = GF(4^3)$. Let β be a primitive element of $GF(q^3)$. Then $GF(q^3) = GF(2^6) = \{0, 1, \beta, \beta^2, \cdots, \beta^{61}, \beta^{62}\}$. Suppose $\alpha = \beta^{21}$, then $GF(q) = GF(4) = \{0, 1, \alpha, \alpha^2\}$. We know that $[GF(q^3) : GF(q)] = 3$, $GF(q^3)$ is a 3-dimensional vector space over $GF(q)$. We can prove that for any $a_0, a_1, a_2 \in GF(q) = \{0, 1, \alpha, \alpha^2\}$, $a_0 + a_1\beta + a_2\beta^2 = 0$ if and only if $a_0 = a_1 = a_2 = 0$, i.e., $1, \beta, \beta^2$ are linear independent over $GF(q)$. So $1, \beta, \beta^2$ is a basis of $GF(q^3)$ over $GF(q)$. Let x, y, z be variables. Then, $(x + y\beta + z\beta^2)^{q+1} = g_0(x, y, z) + g_1(x, y, z)\beta + g_2(x, y, z)\beta^2$, where

$$\begin{aligned} g_1(x, y, z) &= x^2 + \alpha xy + \alpha^2 y^2 + yz + \alpha xz + z^2, \\ g_2(x, y, z) &= \alpha^2 y^2 + yz + \alpha^2 xz + \alpha yz + \alpha z^2, \\ g_3(x, y, z) &= xz + \alpha xy + \alpha^2 y^2 + \alpha^2 yz + z^2. \end{aligned}$$

Obviously, $g_1(x,y,z), g_2(x,y,z)$ and $g_3(x,y,z)$ are polynomials with three variables and coefficients in $GF(q)$. On the other hand, since

$$((x + y\beta + \beta^2)^{q^2+q+1})^q = (x + y\beta + \beta^2)^{q^2+q+1},$$

$(x + y\beta + \beta^2)^{q^2+q+1}$ is also a polynomial with coefficients in $GF(q)$.

Now let $\mathbf{h}_1 = 1, \mathbf{h}_2 = x, \mathbf{h}_3 = y, \mathbf{h}_4 = z, \mathbf{h}_5 = g_0(x,y,z), \mathbf{h}_6 = g_1(x,y,z), \mathbf{h}_7 = g_3(x,y,z), \mathbf{h}_8 = (x + y\beta + \beta^2)^{q^2+q+1}$. For any polynomial $\mathbf{h}_i$,we define its evaluated vector as $\mathbf{h}_i = (\mathbf{h}_i(P_1), \mathbf{h}_i(P_2), \cdots, \mathbf{h}_i(P_{64}))^T$, where $P_1, P_2, \cdots, P_{64}$ are the 64 points in $GF(q^3)$, when $GF(q^3)$ is considered as a three-dimensional vector space over $GF(q)$. Let $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_8]^T$. Then we have a code with parameters:

$$n = q^3 = 64, k = 56.$$

By the following theorem, we know that C is an (64,56, $\geq$5) code.

Theorem 4.2 *Let* $\mathbf{H} = [1, x, y, z, g_0(x,y,z), g_1(x,y,z), g_3(x,y,z), (x + y\beta + \beta^2)^{q^2+q+1}]^T$. *Then* $D_5^{(8)} \leq 3$.

Remark 4.1: In Theorem 7 in [13], Dumer constructed a class of codes over $GF(q)$, where q is a power of an odd prime. The minimum distances of these codes are greater than or equal to 5. According to Dumer, when $n = q^3$, the codes have the following parameters:

$$n = q^3, r = 8, d \geq 5.$$

These codes are known to be optimal in the sense that no other codes with $d \geq 5$ and the sane code lengths have fewer number of parity checks. But, unfortunately, these codes are defined only on $GF(q)$ for odd prime number q. Our code C has the same parameters and is defined on $GF(2^3)$.

5 Conclusions

Bezout's theorem was used to determine an upper bounds of the numbers of common points of two-dimensional polynomials in [1]. With the upper bounds, lower bounds of the minimum distance and generalized Hamming weights of linear codes were also given in [1]. In this paper, we generalize the results in [1] to n-dimensional spaces. we not only give the upper bounds on the number of the intersection points of n-dimensional polynomials and the lower bounds on the generalized Hamming weights, but also give the exact values of the generalized Hamming weights $d_h(C_r)$ for large enough h. With the results in this paper, new methods for constructing more efficient linear codes can be built. They will be applied in computer memory systems, distributed systems, CD audio, Video disk, and CD ROM.

References

[1] Gui-Liang Feng, T. R. N. Rao, Gene A. Berg and Junmei Zhu "Generalized Bezout's Theorem and Its Applications in Coding Theory", to appear in *IEEE Trans. on Information Theory*, Nov., 1997.

[2] Gui-Liang Feng, T. R. N. Rao, "A Class of Algebraic Geometric Codes from Curves in High-Dimensional Projective Spaces", Lecture Notes in Computer Science 673, Springer-Verlag, pp. 132-146, 1993.

[3] R. E. Blahut, *Theory and Practice of Error Control Codes*, Reading, Ma: Addison-Wesley, 1985.

[4] V. K. Wei, "Generalized Hamming weights for linear codes", *IEEE Trans. on Information Theory*, Vol. IT-37, pp. 1412-1428, Sept., 1991.

[5] J. W. P. Hirschfeld, M. A. Tsfasman, and S. G. Vladut, "The Weight Hierarchy of Higher Dimensional Hermitian Codes," *IEEE Trans. on Information Theory*, Vol. 40, pp. 275-278, January 1994.

[6] K. Yang, P. V. Kumar, and H. Stichtenoth, "On the Weight Hierarchy of Geometric Goppa Codes," *IEEE Trans. on Information Theory*, Vol. IT-40, pp. 913-920, May 1994.

[7] S. S. Abhyankar, *Algebraic Geometry for Scientists and Engineers*, The American Mathematical Society, 1990.

[8] J.J. Sylvester, "On a general method of determining by mere inspection the derivations from two equations of any degree", *Philosophical Magazine*, 16(1840), 132-135.

[9] G. L. Feng and T. R. N. Rao, "A Simple Approach for Construction of Algebraic Geometric Codes from Affine Plane Curves," *IEEE Trans. on Information Theory*, Vol. IT-40, No.4, pp. 1003-1012, July 1994.

[10] G. L. Feng and T. R. N. Rao, "Improved Geometric Goppa Codes, Part I: Basic Theory" to appear in *IEEE Trans. on Information Theory.*

[11] S. Miura and N. Kamiya, "Geometric-Goppa Codes on Some Maximal Curves and Their Minimum Distance," *Proceedings of 1993 IEEE Information Theory Workshop*, June 4-8, 1993 at Shizuoka, Japan, pp. 85-86.

[12] Carlos Munuera, "On the Generalized Hamming Weights of Geometric Goppa Codes," *IEEE Trans. on Information Theory*, Vol. 40, No. 6, pp. 2092-2099, 1994.

[13] I. Dumer, "Nonbinary double-error-correcting codes designed by means of algebraic varieties," *IEEE Trans. on Information Theory*, Vol. IT-41, No. 6, PP. 1657-1666, 1995.

Decoding Reed Solomon Codes beyond the Error-Correction Diameter

MADHU SUDAN
Laboratory for Computer Science, MIT
madhu@lcs.mit.edu

Abstract

We describe a new algorithm for the decoding of Reed Solomon codes. This algorithm was originally described in [12]. While the algorithm presented in this article is the same, the presentation is somewhat different. In particular we derive more precise bounds on the performance of the algorithm and show the following: For an $[n, \kappa n, (1-\kappa)n]_q$ Reed Solomon code, the algorithm in [12] corrects $(\epsilon(\kappa) - o(1))n$ errors in polynomial time, where

$$\epsilon(\kappa) = 1 - \frac{1}{1+\rho_\kappa} - \frac{\rho_k}{2}\kappa \text{ where } \rho_\kappa = \left\lfloor \sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{1}{2} \right\rfloor.$$

We also present the following two (hopefully) useful lower bounds on $\epsilon(\kappa)$:

$$\forall \kappa \in [0,1] \quad \epsilon(\kappa) \geq 1 - \sqrt{2\kappa + \frac{\kappa^2}{4}} + \frac{\kappa}{2} \geq 1 - \sqrt{2\kappa}.$$

1 Introduction

For integers n, k and q such that a finite field of size q exists, the Reed Solomon codes are $[n, k, d = n-k]$ codes over the alphabet $F = \mathrm{GF}(q)$ (the Galois field of order q). The code may be obtained by letting a message $m = m_0 \ldots m_{k-1}$ denote the coefficients of a degree $k-1$ polynomial $M(x) = \sum_{j=0}^{k-1} m_j x^j$ and letting the encoding of m be $C(m) = c_1 \ldots c_n$ where $c_i = M(x_i)$ where $x_1, \ldots, x_n$ are n distinct elements of F. (It is standard to pick $n = q-1$, though not necessary to prove the distance property or for the algorithms we describe).

In this paper we consider the following bounded distance decoding problem:

`Input`: Integers n, k and e; and n pairs $\{(x_i, y_i)\}_{i=1}^n$, $x_i, y_i \in F$ with x_i's being distinct[1].
`Goal`: Find all polynomials $p_1, \ldots, p_m$ of degree $k-1$ such that for every $j \in [m]$, $p_j(x_i) \neq y_i$ for at most e values of $i \in [n]$.

---

[1] For our algorithm from [12] we can replace this condition with the weaker one that (x_i, y_i)'s are distinct.

As can be seen easily, this problem captures the bounded distance decoding problem for Reed Solomon codes. There is a rich history of work associated with this problem. The classical work of Berlekamp-Massey (cf. [2, 9]), corrects upto $\lfloor (d-1)/2 \rfloor$ errors. Sidelnikov [11] and Dumer [4] have constructed algorithms which correct up to $\lfloor (d-1)/2 \rfloor + c \log n$ errors for any constant c [4, 11]. We give an algorithm that improves over these results when k/n is sufficiently small (i.e., less than 1/3). Our algorithm is motivated by an algorithm of Welch and Berlekamp [14, 3] which corrects $\lfloor (d-1)/2 \rfloor + 1$ errors. In this article we describe the algorithm of Welch and Berlekamp and use it motivate our decoding algorithm. We also describe a crucial intermediate step from [1] which forms the basis of our algorithm. Our main result is summarized below and proven in Lemma 9.

Theorem 1 *For every ϵ, κ The bounded distance decoding problem with parameters n, $k = \kappa n$ and $e = \epsilon(\kappa) n$ can be solved in polynomial time provided*

$$\epsilon(\kappa) < 1 - \frac{1}{1+\rho_\kappa} - \frac{\rho_\kappa}{2}\kappa \text{ where } \rho_\kappa = \left\lfloor \sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{1}{2} \right\rfloor .$$

2 Decoding with univariate rational functions [14, 3]

The idea of Welch and Berlekamp [14, 3] can be informally described as follows: They describe how a "rational function" in x can be used to "explain" the "data" $\{(x_i, y_i)\}_{i=1}^n$. They then show how to efficiently *find* a rational function that explains the data and then show how to use this rational function to find the (unique) polynomial p which disagrees with the data in at most e places. We now describe the algorithms more formally.

Lemma 2 ([14, 3]) *Given n points $\{(x_i, y_i)\}_{i=1}^n$ such that there exists a degree $k-1$ polynomial p such that $y_i \neq p(x_i)$ for at most e values of i, the following hold:*

1. *There exist polynomials $N(x)$ and $D(x)$ where $\deg(N) \leq k-1+e$, D is monic with $\deg(D) = e$, such that for every $i \in [n]$, $N(x_i) = y_i D(x_i)$. (Informally, we could say that y_i equals the rational function $\frac{N}{D}(x_i)$.)*
2. *Such a pair of polynomials (N, D) can be found in polynomial time.*
3. *For any such pair of polynomials (N, D), $\frac{N}{D}(\cdot) = p(\cdot)$, provided $e \leq \frac{1}{2}(n-k)$.*

Remark: As a consequence p can be found in polynomial time. We just divide the polynomials N and D to obtain p.

Proof: Let $E(x)$ be an "error-locator" polynomial, i.e., $E(x_i) = 0$ if (but not necessarily only if) $y_i \neq p(x_i)$. Notice that E has degree at most e and w.l.o.g. we can allow it to be monic and have degree exactly e. We now notice that the polynomials $D(\cdot) = E(\cdot)$ and $N(\cdot) = p(\cdot)E(\cdot)$ satisfy the condition $N(x_i) = y_i D(x_i)$ for every i. This proves Part 1.

To see Part 2, notice that if we let the unknowns $\{n_j\}_{j=0}^{e+k-1}$ denote the coefficients of $N(\cdot)$ and let the unknowns $\{d_j\}_{j=0}^{e}$ denote the coefficients of $D(\cdot)$, then the constraints $N(x_i) = y_i D(x_i)$ give linear constraints on the unknowns $\{n_j\}$'s and $\{d_j\}$'s. Also the constraint $\deg(D) = e$. We do so by setting the linear constraint $d_e = 1$. Thus a solution pair (N, D) can be found efficiently by solving a linear system.[2]

Finally notice that there exists a pair (N, D) such that $\frac{N}{D}(\cdot) = p(\cdot)$, as described in the proof of Part 1. Thus to prove Part 3, it suffices to prove that for any pair of solutions (N_1, D_1) and (N_2, D_2), satisfying $N_j(x_i) = y_i D_j(x_i)$ every $i \in [n]$ and $j \in \{1, 2\}$, $\frac{N_1}{D_1}(\cdot) = \frac{N_2}{D_2}(\cdot)$. To see this, first observe that for every $i \in [n]$, we have $N_1(x_i)D_2(x_i)y_i = N_2(x_i)D_1(x_1)y_i)$. Futhermore, we can cancel y_i from both sides (even if $y_i = 0$, since in such a case we have $N_1(x_i) = N_2(x_i) = 0$. This yields that for every i, $(N_1D_2)(x_i) = (N_2D_1)(x_i)$. But then both sides describe polynomials of degree $2e + k - 1$ and two sides agree on n points. By the condition on n, we have that $n > 2e + k - 1$ and thus the polynomials on the two sides are identical, i.e., $(N_1D_2)(\cdot) = (N_2D_1)(\cdot)$. This yields the desired conclusion immediately. ▌

3 Decoding with algebraic curves in the plane [1, 12]

A slightly different interpretation of the Welch-Berlekamp algorithm is that it finds an algebraic curve in the plane which "explains" the "data". To be precise, the algorithm finds a function $Q(x, y)$, where $Q(x, y) = D(x)y - N(x)$, such that for every $i \in [n]$, $Q(x_i, y_i) = 0$. While this particular scenario attempts to explain the data by a "linear" polynomial in y - there is no need to restrict the analysis to this situation. Ar et al. [1] considered such a generalization. They consider the case where the data is "explained" some algebraic curve Q of low degree in y (but not necessarily a linear polynomial in y). They show that in such a case, if there exists a polynomial p such that $y_i = p(x_i)$ for many values of i (compared to $\deg_x(Q)$ and $\deg_y(Q)$) then p can be reconstructed easily. To describe their analysis, the following definition is useful.

Definition 3 *For positive integers w_x and w_y, the (w_x, w_y)-weighted degree of a bivariate polynomial $Q(x, y) = \sum_{i,j} q_{ij}x^iy^j$ is defined to be* $\max\{iw_x + jw_y | q_{ij} \neq 0\}$.

Lemma 4 ([1]) *Given n points $\{(x_i, y_i)\}_{i=1}^n$ s.t. there exists a bivariate polynomial Q satisfying:*

The $(1, k-1)$ weighted degree of Q is at most D, $Q \not\equiv 0$ and $\forall i \in [n]$, $Q(x_i, y_i) = 0$. (1)

Then the following hold:

1. *A polynomial Q satisfying equation (1) can be found in polynomial time.*

---

[2] Actually Berlekamp and Welch [14, 3] give a much more efficient solution for this task, but we will not describe their solution here.

2. *If p is a polynomial in x of degree at most $k-1$ such that $y_i \neq p(x_i)$ for at most $e < n - D$ values of i then for any polynomial Q satisfying (1), it is the case that $(y - p(x))$ divides $Q(x,y)$.*

Remark: As a consequence, a small set of polynomials which includes p can be found in polynomial time. We simply factor the polynomial Q obtained in Part 1 above and output all p such that $y - p(x)$ divides Q. The polynomial Q can be factored in time polynomial in its degree using the algorithm of Kaltofen [7] or Grigoriev [6] (see also Kaltofen [8]).

Proof: For Part 1, we observe as in the proof of Lemma 2 that for any i, the condition $Q(x_i, y_i) = \sum_{jl} q_{jl} x^j y^l = 0$ is a linear constraint on the unknowns q_{jl}. Thus a solution satisfying (1) can be found in polynomial time, if one exists.

For Part 2, we consider the polynomial $g(x) \overset{\text{def}}{=} Q(x, p(x))$. Notice that since the $(1, k-1)$-weighted degree of Q is D, the degree of g is also at most D. Notice further that if for some $i \in [n]$, $y_i = p(x_i)$, then $g(x_i) = Q(x_i, p(x_i)) = Q(x_i, y_i) = 0$. Thus g is zero on $n - e > D$ points. Thus g is identically 0. Now consider the polynomial $Q_x(y) \overset{\text{def}}{=} Q(x,y)$ (i.e., the polynomial Q viewed as a polynomial in y with coefficients from the ring of polynomials in x). By the division theorem for polynomials we have that if $Q_x(\zeta) = 0$, then $y - \zeta$ divides $Q_x(y)$. Applying this fact to $Q_x(p(x)) = Q(x, p(x)) = g(x) = 0$ we find that $y - p(x)$ divides $Q(x,y)$. ▮

Notice the close correspondence between Lemmas 2 and 4. The main difference between the two is Part 1 of Lemma 2, the analog of which is missing in Lemma 4. As a result Lemma 4 works conditionally, i.e., only when the "data" is explained by a low-degree algebraic curve. Our solution complements this by providing a "triviality" result. Namely, we observe that every set of points lies on a low degree algebraic curve; low enough to make Lemma 4 always useful! We illustrate this observation by a simple example

Example 5 *For any set of points $\{x_i, y_i\}_{i=1}^n$ there exists a non-zero polynomial Q with* $\deg_x(Q), \deg_y(Q) \leq \lceil \sqrt{n} \rceil$.

Proof: Consider the linear system $Q(x,y) = \sum_{j=0}^{\lceil \sqrt{n} \rceil} \sum_{l=0}^{\lceil \sqrt{n} \rceil} q_{jl} x_i^j y_i^l = 0$. For every i this forms a homogenous linear system. The system has n constraints on $(\lceil \sqrt{n} + 1 \rceil)^2 \geq (\sqrt{n} + 1)^2 \geq n + 1$ unknowns. By just counting the number of unknowns we know that the system has a non-zero solution (any homogenous linear system with more variables than unknowns has a non-zero solution). ▮

Notice that unlike in Part 1 of Lemma 2, we didn't even use the fact that the data has some agreement with a degree $k-1$ polynomial. Thus the example above is a "triviality" result - it holds for any set of n points. Yet the trivial result can be useful as seen next:

Example 6 *For any set of n distinct points $\{(x_i, y_i)\}_{i=1}^n$ the following hold:*

1. *There exists a bivariate polynomial Q satisfying equation (1) with $D = k\sqrt{n}$.*
2. *Such a polynomial Q can be found in polynomial time.*
3. *If p is a polynomial in x of degree at most $k-1$ such that $y_i \neq p(x_i)$ for at most $e < n - k\lceil\sqrt{n}\rceil$ values of i then for any polynomial Q satisfying (1) with $D = k\sqrt{n}$, it is the case that $(y - p(x))$ divides $Q(x, y)$.*

Proof: Part 1 follows from Example 5. Parts 2 and 3 follow from Parts 1 and 2 of Lemma 4 with $D = k\lceil\sqrt{n}\rceil$. ▌

Notice that Example 6 is already correcting more errors than guaranteed by Lemma 2 for some values of n and k (in particular when k grows as $o(\sqrt{n})$. By some fine tuning we can actually get to the point where this algorithm always does at least as well as the Welch-Berlekamp algorithm and for rate less than 1/3 it starts correcting significantly more error. The fine tuning is performed by minimizing the weighted $(1, k-1)$ degree of the polynomial Q which is used to explain the data.

Lemma 7 *Given n and k, let t be the smallest positive integer satisfying:*

$$(t-1)\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor+1\right)-\frac{(k-1)}{2}\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor+1\right)\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor\right)>n.$$

Then, for any set of n distinct points $\{(x_i, y_i)\}_{i=1}^n$ the following hold:

1. *There exists a bivariate polynomial Q satisfying:*

   The $(1, k-1)$ weighted degree of Q is $\leq t$, $Q \not\equiv 0$ and $\forall i \in [n]$, $Q(x_i, y_i) = 0$. (2)

2. *A polynomial Q satisfying (2) can be found in polynomial time.*
3. *If p is a polynomial in x of degree at most $k-1$ such that $y_i \neq p(x_i)$ for at most $e < n-t$ values of i then for any polynomial Q satisfying (2), it is the case that $(y-p(x))$ divides $Q(x, y)$.*

Remark: Again as a consequence we can find a list of all polynomials p that agree with the points $\{(x_i, y_i)\}$ in $n - e$ points in time polynomial in n.

Proof: A polynomial Q of $(1, k-1)$-weighted degree t is allowed to have a non-zero coefficient q_{jl} if $j + (k-1)l \leq t$. Counting the number of such coefficients we find that Q has

$$\sum_{l=0}^{\lfloor\frac{t-1}{k-1}\rfloor}(t-(k-1)l+1)=(t-1)\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor+1\right)-\frac{(k-1)}{2}\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor+1\right)\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor\right)$$

coefficients. Since this number is strictly greater than n we must have a non-zero solution. This proves Part 1. Parts 2 and 3 follow directly from Lemma 4. ▌

4 Some bounds

Let $e(k,n)$ denote the maximum number of errors corrected by the algorithm of Lemma 7 on an $[n,k,d]_q$ Reed Solomon code. Let $\kappa = k/n$ denote the rate of a code and let $\epsilon(\kappa) \triangleq \lim_{n\to\infty} e(\kappa n, n)/n$ denote the asymptotic error-correcting rate of the algorithm. In this section we derive the exact form of $e(k,n)$ and $\epsilon(\kappa)$. We also describe some lower bounds to both quantities.

Lemma 8

$$e(k,n) = n-(k-1)r_{k,n} - \left\lceil \frac{2n-(k-1)r_{k,n}(r_{k,n}+1)}{2(r_{k,n}+1)} \right\rceil - 2 \textit{ where } r_{k,n} = \left\lfloor \sqrt{\frac{2(n+1)}{k-1}+\frac{1}{4}} - \frac{1}{2} \right\rfloor.$$

Proof: Let t be the smallest integer satisfying

$$(t-1)\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor+1\right) - \frac{(k-1)}{2}\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor+1\right)\left(\left\lfloor\frac{t-1}{k-1}\right\rfloor\right) > n.$$

By Lemma 7 $e(k,n) = n-t-1$. Let $t-1 = r(k-1)+s$, where r and s are integers and $0 \le s < k-1$. Then $\lfloor\frac{t-1}{k-1}\rfloor = r$; and r and s satisfy

$$(r(k-1)+s)(r+1) - \frac{(k-1)(r+1)r}{2} > n. \tag{3}$$

We first determine r based on the above. The constraints $0 \le s < k-1$ translate into the constraints:

$r(k-1)(r+1) - r(k-1)(r+1)/2 \le n+1$ and $(r+1)(k-1)(r+1) - r(k-1)(r+1)/2 > n+1$.

The above, in turn, simplify to

$$r \le \sqrt{\frac{2(n+1)}{k-1} - \frac{1}{4}} - \frac{1}{2} \text{ and } r > \sqrt{\frac{2(n+1)}{k-1}} - 1 - 1.$$

The requirement that r be an integer now allows us to determine r precisely:

$$r = \left\lfloor \sqrt{\frac{2(n+1)}{k-1}+\frac{1}{4}} - \frac{1}{2} \right\rfloor.$$

The value s can now be determined from (3) and we get:

$$s = \left\lceil \frac{2(n+1) - r(r+1)(k-1)}{2(r+1)} \right\rceil.$$

The lemma follows by setting $r_{k,n} = r$ and $t = r(k-1)+s+1$ and using the fact that $e(k,n) = n-t-1$. ▮

The following lemma simplifies some of the expressions above by examining the error-correction rate $\epsilon(\kappa)$ as a function of κ.

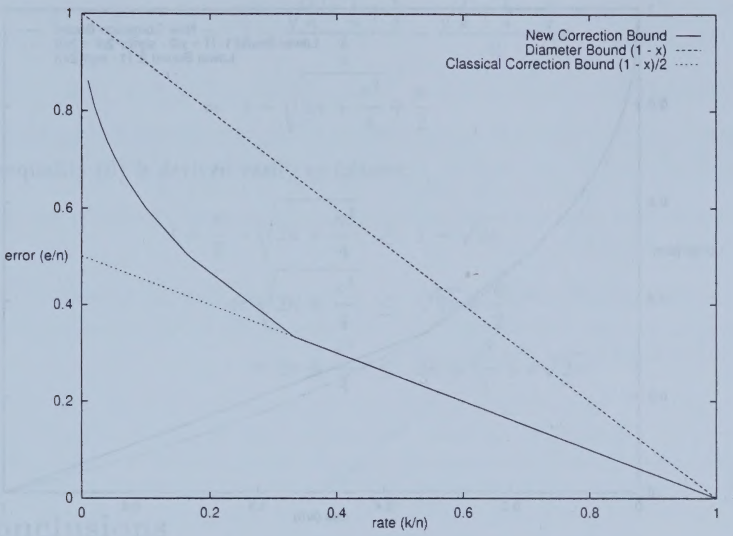


Figure 1: The new error correction bound $\epsilon(\kappa)$ plotted against the rate κ. Also described are the error correction radius of the code and the distance of the code.

Lemma 9

$$\epsilon(\kappa) = 1 - \frac{1}{1+\rho_\kappa} - \frac{\rho_\kappa}{2}\kappa \textit{ where } \rho_\kappa = \left\lfloor \sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{1}{2} \right\rfloor. \tag{4}$$

$\epsilon(\kappa)$ can be lower bounded as follows:

$$\begin{aligned} \forall \kappa \in [0,1], \quad \epsilon(\kappa) &\geq 1 - \sqrt{2\kappa + \frac{\kappa^2}{4}} + \kappa/2 & (5) \\ &\geq 1 - \sqrt{2\kappa} & (6) \end{aligned}$$

Remark: The bound (4) is described pictorially in Figure 1. The lower bounds (5) and (6) are compared against the bound (4) in Figure 2.

Proof: (4) follows from Lemma 8 by letting $\rho_\kappa = \lim_{n\to\infty} r_{\kappa n,n}$ and simplifying the quantity $\lim_{n\to\infty} e(\kappa n, n)/n$.

To prove (5), we first show that

$$1 - \frac{1}{1+\rho_\kappa} - \frac{\rho_\kappa}{2}\kappa \geq 1 - \frac{1}{1+\rho'_\kappa} - \frac{\rho'_\kappa}{2}\kappa, \tag{7}$$

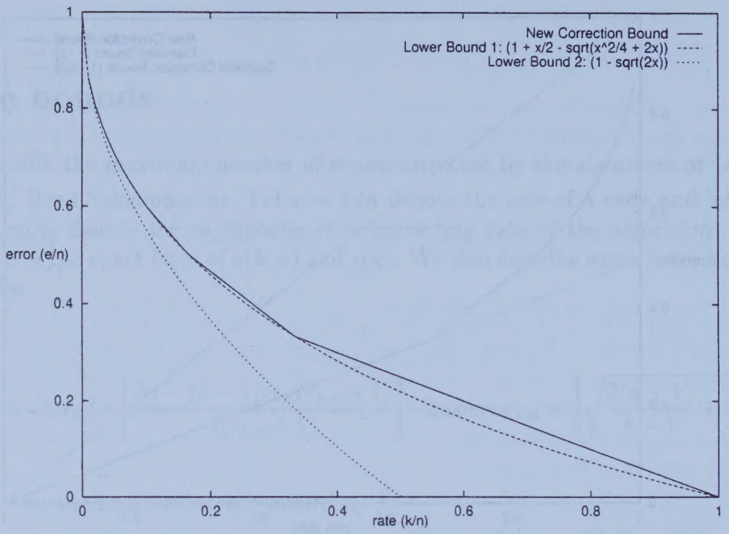


Figure 2: Lower bounds on the error-correction rate of our algorithm.

where $\rho'_\kappa = \sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{3}{2}$. Notice that by the definition of ρ'_κ, we have

$$\rho'_\kappa \leq \rho_\kappa \leq \rho'_\kappa + 1 \text{ and } \frac{2}{\kappa} = (\rho'_\kappa + 2)(\rho'_\kappa + 1).$$

(7) follows from:

$$\begin{array}{rrcll} & \frac{1}{1+\rho'_\kappa} - \frac{1}{1+\rho_\kappa} + \frac{\rho'_\kappa}{2}\kappa - \frac{\rho_\kappa}{2}\kappa & \geq & 0 & \\ \Leftarrow & (\rho_\kappa - \rho'_\kappa)\left(\frac{1}{(1+\rho'_\kappa)(1+\rho_\kappa)} - \frac{\kappa}{2}\right) & \geq & 0 & \\ \Leftarrow & \frac{1}{(1+\rho'_\kappa)(1+\rho_\kappa)} - \frac{\kappa}{2} & \geq & 0 & (\text{Since } \rho_\kappa \geq \rho'_\kappa) \\ \Leftarrow & \frac{1}{(1+\rho'_\kappa)(1+\rho_\kappa)} & \geq & \frac{1}{(1+\rho'_\kappa)(2+\rho'_\kappa)} & (\text{Since } \frac{2}{\kappa} = (1+\rho'_\kappa)(2+\rho'_\kappa)) \\ \Leftarrow & 2 + \rho'_\kappa & \geq & 1 + \rho_\kappa. & \end{array}$$

where the last inequality follows from the properties of ρ'_κ.

(5) now follows from the following equalities:

$$1 - \frac{1}{1 + \sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{3}{2}} - \frac{\sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{3}{2}}{2}\kappa$$

$$= 1 - \frac{1}{\sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{1}{2}} - \frac{\sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{3}{2}}{2}\kappa$$

$$= 1 - \frac{\sqrt{\frac{2}{\kappa} + \frac{1}{4}} + \frac{1}{2}}{\frac{2}{\kappa}} - \frac{\sqrt{\frac{2}{\kappa} + \frac{1}{4}} - \frac{3}{2}}{2}\kappa$$

$$= 1 - \sqrt{2\kappa + \frac{\kappa^2}{4}} + \frac{\kappa}{2}.$$

Finally inequality (6) is derived easily as follows:

$$\begin{aligned} 1 + \frac{\kappa}{2} - \sqrt{2\kappa + \frac{\kappa^2}{4}} &\geq 1 - \sqrt{2\kappa} \\ \Leftarrow \sqrt{2\kappa + \frac{\kappa^2}{4}} &\leq \sqrt{2\kappa} + \frac{\kappa}{2} \\ \Leftarrow 2\kappa + \frac{\kappa^2}{4} &\leq 2\kappa + \frac{\kappa^2}{4} + \kappa\sqrt{2\kappa} \end{aligned}$$

5 Conclusions

The algorithm as presented here does generalize naturally to dealing with erasures, for the simple reason that we made no assumptions about the x_i's (we didn't depend on them being in some specific order etc.). However the bounds can not be represented elegantly in any form - so we do not attempt to do so here. We only point out that since the Welch-Berlekamp algorithm is a special case of ours, we do at least as well.

There are a number of questions on decoding Reed Solomon codes that remain open. For instance, is there an algorithm that can decode from more errors (than $(d-1)/2$) when $\kappa = k/n > 1/3$? A nice target would be a decoding algorithm that works for $\epsilon(\kappa) \leq 1 - \sqrt{\kappa}$. In this case we know (cf. [5, 10]) that the number of codewords within a distance of $\epsilon(\kappa)n$ is bounded by a polynomial in n. One does expect that the problem will become harder as $\epsilon(\kappa) \to 1 - \kappa$. It would be interesting to see if the problem becomes NP-hard as $\epsilon(\kappa) \to 1 - \kappa$.

Acknowledgments

I am grateful to Dave Forney, Simon Litsyn, Ronny Roth, and Alex Vardy for their valuable comments on the article [12]. Their detailed comments motivated and encouraged the writing of this article.

References

[1] S. Ar, R. Lipton, R. Rubinfeld and M. Sudan. Reconstructing algebraic functions from mixed data. *SIAM Journal on Computing*, to appear. Preliminary version in

Proceedings of the 33rd Annual IEEE Symposium on Foundations of Computer Science, pp. 503–512, 1992.

[2] E. R. BERLEKAMP. *Algebraic Coding Theory.* McGraw Hill, New York, 1968.

[3] E. R. BERLEKAMP. Bounded Distance +1 Soft-Decision Reed-Solomon Decoding. In *IEEE Transactions on Information Theory*, pages 704-720, vol. 42, no. 3, May 1996.

[4] I. I. DUMER. Two algorithms for the decoding of linear codes. *Problems of Information Transmission*, 25(1):24–32, 1989.

[5] O. GOLDREICH, R. RUBINFELD AND M. SUDAN. Learning polynomials with queries: The highly noisy case. *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*, pp. 294–303, 1995.

[6] D. GRIGORIEV. Factorization of Polynomials over a Finite Field and the Solution of Systems of Algebraic Equations. Translated from Zapiski Nauchnykh Seminarov Lenningradskogo Otdeleniya Matematicheskogo Instituta im. V. A. Steklova AN SSSR, Vol. 137, pp. 20-79, 1984.

[7] E. KALTOFEN. A Polynomial-Time Reduction from Bivariate to Univariate Integral Polynomial Factorization. In *23rd Annual Symposium on Foundations of Computer Science*, pages 57-64, 1982.

[8] E. KALTOFEN. Polynomial factorization 1987–1991. *LATIN '92*, I. Simon (Ed.) Springer LNCS, v. 583:294–313, 1992.

[9] F. J. MACWILLIAMS AND N. J. A. SLOANE. *The Theory of Error-Correcting Codes.* North-Holland, Amsterdam, 1981.

[10] J. RADHAKRISHNAN. *Personal communication*, January, 1996.

[11] V. M. SIDELNIKOV. Decoding Reed Solomon codes beyond $(d-1)/2$ and zeros of multivariate polynomials. *Problems of Information Transmission*, 30(1):44–59, 1994.

[12] M. SUDAN. Decoding of Reed Solomon codes beyond the error-correction bound. *Journal of Complexity*, 13(1):180-193, March 1997.

[13] A. VARDY. Algorithmic complexity in coding theory and the minimum distance problem. *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pp. 92–109, 1997.

[14] L. R. WELCH AND E. R. BERLEKAMP. Error correction of algebraic block codes. *US Patent* Number 4,633,470, issued December 1986.

Decoding Algebraic-Geometric Codes beyond the Error-Correction Bound

M. Amin Shokrollahi* Hal Wasserman†

Extended Abstract

Following a construction of Goppa [7], one can use algebraic curves over finite fields to design linear error-correcting codes called *algebraic-geometric codes* or *AG-codes.* These codes can be viewed as a generalization of Reed-Solomon codes, for an AG-codeword is formed by evaluating at specified rational points a function in the function field of a curve. Given a curve of genus g, n rational points, and a specified *designed distance* $d^* \leq n - g$, one can construct an AG-code which is assured to have minimum distance at least d^*.

The encoding algorithm for an AG-code is simple once one knows the n rational points and a basis for the functions to be evaluated. However, corresponding decoding algorithms are more complicated. Several such algorithms have recently appeared [2, 3, 4, 5, 8, 9, 13, 14, 15, 16, 17, 20]. These algorithms decode up to $\lfloor (d^* - 1 - \ell)/2 \rfloor$ errors, where ℓ is some integer $0 \leq \ell \leq g$ and g is the genus of the curve. They perform between $O(n^2)$ and $O(n^3)$ operations over the base field $\mathbf{F}_q$.

Most of these algorithms fail if the number of errors exceeds the designed error-correction bound $(d^* - 1)/2$, and all of them fail if the number of errors is larger than the true error-correction bound $(d-1)/2$, where d is the minimum distance of the code. Beyond the true error-correction bound it is in general not possible to uniquely reconstruct a codeword from a received word. However, one might hope to reconstruct a small set of candidates which provably includes the original codeword. This motivates the following definition.

Definition. *An* $[n, k, e, b]_q$*-code is an* $[n, k]_q$*-code such that the Hamming sphere of radius* e *around any* $y \in \mathbf{F}_{q^n}$ *contains at most* b *codewords.*

For example, note that an $[n, k, 2e + 1]_q$-code is an $[n, k, e, 1]_q$-code.

The decoding task associated with an $[n, k, e, b]_q$-code requires methods quite different from those customarily employed for error-correction. This task was first considered by Sudan [18, 19], who investigated alternative decoding algorithms for Reed-Solomon codes. He derived the surprising result that an $[n, k]_q$ Reed-Solomon code is an $[n, k, e, b]_q$-code such that e is approximately $n - \sqrt{2kn}$ and b is approximately $\sqrt{2n/k}$. Hence e can be close to n while k/n is bounded away from zero and b is small. Moreover, Sudan

---

[*] International Computer Science Institute, Berkeley, CA. Research supported by a Habilitationsstipendium of the Deutsche Forschungsgemeinschaft, Grant Sh 57/1-1.

[†] Computer Science Division, University of California, Berkeley, CA. Supported by a Department of Education GAANN Fellowship.

designed a corresponding polynomial-time decoding algorithm based on factorization of bivariate polynomials. Sudan's algorithm thus has an unprecedented ability to efficiently reconstruct meaningful data in spite of very high amounts of noise. Reed-Solomon codes are however of limited utility, as they require the block length n to be no larger than the field size q.

Here we generalize Sudan's results to the class of AG-codes. Our main result is as follows:

Main Theorem. *Let C be an $[n,k,d]_q$-AG-code built over an algebraic function field K of genus g. Let $\alpha := k+g-1$ and $\beta := \lceil\sqrt{2\alpha n}+g-1\rceil$. Then C is an $[n,k,n-\beta-1,\lceil\sqrt{2n/\alpha}\,\rceil]_q$-code.*

We also specify a corresponding decoding algorithm, which runs in time polynomial in n and in the size of the basis functions used for encoding, where a function's *size* is a bound on the complexity of evaluating it at a rational point. As for Sudan's method, the crucial tool for our general algorithm is polynomial factorization. However, while Sudan needs to factor a bivariate polynomial over a finite field, we must investigate the more general problem of factoring a univariate polynomial over an algebraic function field. Our algorithm for this task is an adaptation of a well-known procedure for factoring polynomials over number fields, as described in, e.g., [1, 10, 11].

Applying our Main Theorem to the asymptotically good sequence of Garcia-Stichtenoth function fields [6] allows us to construct sequences of $[n,k,e,b]_{q^2}$-codes having excellent parameters.

The bottleneck of our algorithm is the factorization routine. Therefore, we derive for small b variants of our general algorithm in which the factorization step has been eliminated. One variant, which corresponds to the case of decoding $[n,k,e,1]_q$-codes (i.e., conventional decoding), is particularly interesting. It is slightly less powerful than existing AG-decoding algorithms, as it can correct only up to $(d^*-1)/2-g$ errors. However, its implementation only requires solving an $n\times(n+1)$ system of linear equations. Another variant corresponds to the case of decoding $[n,k,e,2]_q$-codes. We find that the small increase from $b=1$ to $b=2$ leads to a substantial result, allowing us to reconstruct after a number of errors which may approach $2n/3$. Moreover, the associated decoding algorithm is again found to be of interest because of its enhanced efficiency. One of the ingredients of this algorithm is a novel procedure for solving quadratic equations over finite fields of arbitrary characteristic.

A natural property of the "highly noisy" decoding algorithms discussed above is that they may not return a *unique* solution. We argue that this does not limit the utility of such algorithms, as they can potentially be combined with other strategies, such as soft decision decoding, to identify the one true solution. Moreover, we show that, if a large fraction of the letters of a codeword are corrupted by noise which is random in a specified sense, it may be possible to *uniquely* decode. This inspires us to speak of an $[n,k,e,1]_q^p$-code, which is defined to be an $[n,k]_q$-code such that, with probability of failure $\le p$, a codeword may be uniquely reconstructed after up to e of its letters have been replaced with uniform-random letters. We will demonstrate that there exist constant-rate $[n,k,e,1]_q^p$-AG-codes for which e is surprisingly close to n while p is very small. We also specify a corresponding decoding algorithm, which does not require polynomial factorization and hence is particularly efficient.

We also investigate asymptotic properties of $[n,k,e,b]_q$-codes. We define analogues of the function α_q [12, Def. 5.1.2] for our new classes of $[n,k,e,b]_q$-codes and $[n,k,e,1]_q^p$-

codes, and derive lower bounds for these functions by creating asymptotically good sequences of AG-codes.

References

[1] H. Cohen. *A Course in Computational Algebraic Number Theory*, volume 138 of *Graduate Texts in Mathematics.* Springer Verlag, 1993.

[2] C. Dahl. Fast decoding of codes from algebraic curves. *IEEE Trans. Inform. Theory*, 40:223–229, 1994.

[3] I. Duursma. Algebraic decoding using special divisors. *IEEE Trans. Inform. Theory*, 39:694–698, 1993.

[4] D. Ehrhard. Achieving the designed error capacity in decoding algebraic-geometric codes. *IEEE Trans. Inform. Theory*, 39:743–751, 1993.

[5] G.L. Feng and T.R.N. Rao. Decoding algebraic-geometric codes up to the designed minimum distance. *IEEE Trans. Inform. Theory*, 39:37–45, 1993.

[6] A. Garcia and H. Stichtenoth. A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound. *Invent. Math.*, 121:211–222, 1995.

[7] V.D. Goppa. Codes on algebraic curves. *Sov. Math. Dokl.*, 24:170–172, 1981.

[8] J. Justesen, K.J. Larsen, A. Havemose, H.E. Jensen, and T. Høholdt. Construction and decoding of a class of algebraic geometry codes. *IEEE Trans. Inform. Theory*, 35:811–821, 1989.

[9] J. Justesen, K.J. Larsen, H.E. Jensen, and T. Høholdt. Fast decoding of codes from algebraic plane curves. *IEEE Trans. Inform. Theory*, 38:111–119, 1992.

[10] S. Landau. Factoring polynomials over algebraic number fields. *SIAM J. Comput.*, 14:184–195, 1985.

[11] A.K. Lenstra. Factoring polynomials over algebraic number fields. Technical Report IW213/82, Dept. Comp. Science, Stichting Mathematisch Centrum, Amsterdam, 1982.

[12] J.H. van Lint. *Introduction to Coding Theory.* GTM. Springer Verlag, 1982.

[13] M.E. O'Sullivan. Decoding of codes defined by a single point on a curve. *IEEE Trans. Inform. Theory*, 41:1709–1719, 1995.

[14] R. Pellikaan. On a decoding algorithm for codes on maximal curves. *IEEE Trans. Inform. Theory*, 35:1228–1232, 1989.

[15] S.C. Porter, B.Z. Shen, and R. Pellikaan. Decoding geometric Goppa codes using an extra place. *IEEE Trans. Inform. Theory*, 38:1663–1676, 1992.

[16] S. Sakata, J. Justesen, Y. Madelung, H.E. Jensen, and T. Høholdt. Fast decoding of algebraic-geometric codes up to the designed minimum distance. *IEEE Trans. Inform. Theory*, 41:1672–1677, 1995.

[17] A.N. Skorbogatov and S.G. Vladut. On decoding of algebraic-geometric codes. *IEEE Trans. Inform. Theory*, 36:1051–1060, 1990.

[18] M. Sudan. Maximum likelihood decoding of Reed-Solomon codes. In *Proc. 37th FOCS*, pages 164–172, 1996.

[19] M. Sudan. Decoding of Reed-Solomon codes beyond the error-correction bound. *J. Compl.*, 13:180–193, 1997.

[20] S.G. Vladut. On the decoding of algebraic-geometric codes over $\mathbf{F}_q$ for $q \geq 16$. *IEEE Trans. Inform. Theory*, 36:1461–1463, 1990.

Fast Erasure and Error Decoding of Evaluation Codes.

Tom Høholdt
Department of Mathematics
Technical University of Denmark, Bldg. 303
DK-2800 Lyngby, Denmark

I. Introduction

There are at least two ways to extend the Berlekamp - Massey algorithm for decoding BCH codes to cover erasures as well. The standard way, where one works with modified syndromes and a non-standard way, when one initialise the algorithm with the erasure polynomial. Recently S. Sakata [1] and Feng and Rao [2] generalised the standard extension of Berlekamp - Massey to one point algebraic geometry codes. This paper generalises the non-standard approach to the evaluation codes defined in [3], which includes the one point algebraic geometry codes, thereby extending the result of [4]. For one point algebraic geometry codes R. Kötter in [5] used a similar idea but his generalisation is based on the modified iterative algorithm of [6].

II. The Codes.

Let R be an $\mathbf{F}_q$ -algebra and

$$\varphi: R \rightarrow \mathbf{F}_q^{\ n}$$

a surjective algebra morphism, where multiplication in $\mathbf{F}_q^{\ n}$ is componentwise, and suppose $\rho: R \rightarrow \mathbf{N}_0 \cup \{-\infty\}$ is an order function (se [3]).

Let $E_\ell = \{\varphi(f) \mid f \in R, \rho(f) \leq \ell\}$ and $C_\ell = E_\ell^{\perp}$. It can be shown that E_ℓ is a $(n, \ell - g + 1, \geq n - \ell)$ code over $\mathbf{F}_q$ and C_ℓ is a $(n, n - \ell + g - 1, \geq \ell - 2g + 2)$ code over $\mathbf{F}_q$, when $3g - 2 < \ell < n - g + 1$ where $g = \#\{\mathbf{N}_0 \cup \{-\infty\} \setminus \rho(R)\}$. For a description of the setup and the decoding in the errors only case see [3].

III. The Erasure- and Error Decoding Algorithm

In the decoding situation we receive a vector $\underline{r}$, which is the sum $\underline{r} = \underline{c} + \underline{e} + \underline{\hat{e}}$ of an unknown codeword $\underline{c}$, an unknown error vector $\underline{e}$ and the erasure vector $\underline{\hat{e}}$ for which we known the nonzero positions but not the values.

We have $S_f(\underline{e} + \underline{\hat{e}}) = \varphi(f) \cdot (\underline{e} + \underline{\hat{e}}) = S_f(\underline{r})$ for all f's with $\rho(f) \leq \ell$. The decoding strategy is from these syndromes to calculate all syndromes $S_f(\underline{e} + \underline{\hat{e}})$, $\rho(f) \leq N$ and from these determine the error values.

A function f has *span* i if $S_{ff_i}(\underline{e} + \underline{\hat{e}}) \neq 0$ but $S_{ff_k}(\underline{e} + \underline{\hat{e}}) = 0$ $k < i$. The Sakata algorithm reads the syndromes in increasing order and has at each step a set F of minimal functions valid so far and a set G of functions which failed to be valid at an earlier stage and which have maximal spans. At each stage of the algorithm let $\Delta_0 = \{j \in N_0 \mid j \leq \text{span}(g), g \in G\}$ and $\Delta_1 = \Delta(F) \setminus \Delta_0$.

In the modification we initialise the set F with the minimal set of erasure functions $F_0 = \{g^{(1)},\ldots, g^{(x)}\}$ and $G = \emptyset$ and start reading the syndromes in increasing order. The aim is to find a minimal set of functions from the ideal $\langle F_0 \rangle$, valid for all syndroms. If a function f from the current F-set fails to be valid at the next order ρ and $\rho - \rho(f) \leq \text{span}(g)$ for some $g \in G$ then f enters the set G and leaves F.

We have allready explained which elements from the current F-set are placed in the new G-set and the new Δ_0-set is changed accordingly.

Now we can prove that

$$\Delta(F_{new}) = \Delta(F_{old}) \cup_f A_f$$

where $A_f = \{j \in N_0 \mid \rho(f) \leq j \leq \rho \text{ and } \rho - j \notin \Delta_0 \text{ old}\}$. This fact means that it is easy to determine Δ_1-new, and then the minimal elements of $N_0 \setminus (\Delta_0 \text{ new} \cup \Delta_1 \text{ new})$ are the orders of the functions in F_{new} and the functions themselves are found by the usual updating procedure. The voting for the unknown syndromes is carried out as before, but the proof must be modified.

REFERENCES

[1] S. Sakata, "Fast Erasures-and-Error Decoding of One-point AG-Codes up to the Feng-Rao Bound", Proceedings of the 1995 IEEE-ISIT, Whistler, Canada.

[2] G.L. Feng, M.S. Kolluro and T.R.N. Rao, "An Erasures-and-Errors Decoding Algorithm for Algebraic-Geometric Codes", Preprint University of Southwestern Louisiana, USA 1994.

[3] T. Høholdt, J.H, van Lint, R. Pellikaan, "Algebraic Geometry Codes in Handbook on Coding Theory (eds. R.A. Brualdi, W.C. Hoffman, V.S. Pless) Elsevier, to appear.

[4] S. Sakata, D.A. Leonard, H.E. Jensen and T. Høholdt, "Fast Erasure and Error Decoding of Algebraic Geometry Codes up to the Feng-Rao Bound. Submitted to IEEE-Trans. Inform. Theory.

[5] R. Kötter, "Fast generalized minimum distance decoding of algebraic-geometric and Reed-Solomon codes", Ph.D. Thesis chapt. 4, Linköping University, Sweden 1996.

[6] G.L. Feng and T.R.N. Rao, "Decoding algebraic geometric codes up to the designed minimum distance", IEEE Trans. Inform. Theory, vol. 39, pp. 36-47, Jan. 1993.

Fast GMD decoding of algebraic codes

Shojiro Sakata
The University of Electro-Communications
Department of Computer Science and Information Mathematics
Chofu-shi, Tokyo 182, JAPAN

Extended Abstract

In this paper we treat GMD decoding of algebraic codes up to their designed distances, where the 'algebraic codes' are a wider class of linear ECC's defined with certain algebraic objects. Such codes are isomorphic to linear subspaces of the coordinate ring R/I, where $R := F[x_1, \cdots, x_N]$ is the N-variate polynomial ring over the finite field F and I is an ideal of R having the Gröbner window $\Delta(I)$ with respect to an admissible total order $\leq_T$ such that $\#\Delta(I) = dim_F R/I = n$ (code length). The class includes most of either theoretically or practically important ECC's, e.g. RS codes, BCH codes, classical Goppa codes, algebraic geometric (AG) Goppa codes.

For those codes, decoding up to the designed distance can be considered to be a one-dimensional (1D) or N-dimensional (ND) shift-register synthesis problem and be executed via BM algorithm (alternatively, Euclidean algorithm) or its extension to the ND case, i.e. BMS algorithm.

More precisely, given the syndrome array $u^q := (u_p | p <_T q)$, $q \in \Delta(I)$, we are required to find a minimal polynomial or a minimal polynomial set of the array u^q defined as follows, where the zeros of such polynomials specify the error locations. In the 1D case of BCH and RS codes, etc., we consider the linear recurrence

$$f[u]_p := \sum_{i=0}^{s} f_i u_{i+p-s} = 0, \; s \leq p < q,$$

corresponding to a polynomial $f = \sum_{i=0}^{s} f_i x^i \in F[x]$ having degree $deg(f) = s$ ($f_s \neq 0$), and the set of minimal polynomials of u^q

$$MinVal(u^q) := \{f \in F[x] | f[u]_p = 0, \; s \leq p < q, \; s(= deg(f)) : \text{minimal}, \; f_s = 1\}.$$

On the other hand, both error-and-erasure decoding and GMD decoding can be handled as constrained shift-register synthesis problems and can be realized via certain modifications of those algorithms, where 'constrained' means that the solution should satisfy some additional constraint derived from the given erasure pattern. The method is based on a single fundamental lemma given by Sakata-Leonard-Jensen-Høholdt (Lemma 4.1 of [1]) or its general version for any dimension and an extension of the theories about BM and BMS algorithms [2][3][4][1].

Again in the 1D case, we are required to find a minimal polynomial or a minimal polynomial set of the errata (=erasures plus errors) syndrome array u^q subject to the additional constraint of having the given erasure positions $\alpha^{\nu_1}, \cdots, \alpha^{\nu_\tau}$

$$MinVal(u^q; I_\tau) := \{f \in I_\tau | f[u]_p = 0, \; s \leq p < q, \; s(= deg(f)) : \text{minimal}, \; f_s = 1\},$$

where $I_\tau := < (x - \alpha^{\nu_1}) \cdots (x - \alpha^{\nu_\tau}) > (\subset F[x])$ is the erasure location ideal. For a polynomial $g \in I_k := < (x - \alpha^{\nu_1}) \cdots (x - \alpha^{\nu_k}) >$, $0 \leq k \leq \tau$, we define the order of g (with respect to u^q) as $ord(g) := p$ such that $g[u]_r = 0$, $t := deg(g) \leq r < p$ and $g[u]_p \neq 0$ and the span of g (with respect to u^q) as $span(g) := ord(g) - deg(g) = p - t$. The following basic lemma and theorems are the main building blocks of our theory and algorithm, where we assume that $f \in MinVal(u^q; I_k)$, $deg(f) = s$, $g \in I_k$, and $span(g) = max\{span(g) | g \in I_k, ord(g) < q\} = c$.

Lemma: Let $ord(f) = q$. Then, there exists $h \in Min(u^{q+1}; I_k)$ such that $deg(h) = s$ if and only if $q - s \leq c$.

Theorem: Let $q - s \leq c$. Then,

$$MinVal(u^q; I_k) = \{f + hg | h \in K[x], deg(h) \leq s - (q - c)\}.$$

Corollary: $f \in MinVal(u^q; I_k)$ is unique, i.e. $\#MinVal(u^q; I_k) = 1$ if and only if $q - s > c$.

The above theorem gives the complete class of all minimal polynomials of u^q and the corollary states the uniqueness condition of a minimal polynomial of u^q. On the basis of those theorems, we can have two alternative ways of finding a minimal polynomial of the errata syndrome array, depending on either pre-processing or post-processing of the erasure locations. Furthermore, there are several choices between both the extremities, i.e. pre-processing and post-processing of erasure locations. Finally, we have a fast GMD decoding algorithm as a modification of BM and BMS algorithm.

Although some part is in principle equivalent to one obtained by Kötter [5] (for AG codes) and the other to each of those by Sorger [6], Taipale et al. [7], Kamiya [8], Kötter [5], etc., respectively (for RS codes), one can get a broader point of view to the soft decision decoding and fast GMD decoding of algebraic codes together with some reduction in decoding complexity and other merits. In particular, the complexity of our method is about half in comparison with any other method including computation of error evaluation polynomials in the iterations of the algorithm. Our method resembles Taipale et al.'s intention, but ours never needs any backtrack procedure which is necessary in their method [7].

References

[1] S. Sakata, D. Leonard, H. Jensen and T. Høholdt "Fast erasure-and-error decoding of AG-codes up to the Feng-Rao bound", submitted to *IEEE Trans. on Inform. Theory.*

[2] J.L. Massey, "Shift-register synthesis and BCH decoding," *IEEE Trans. on Inform. Theory*, vol.15, pp.122–1127, 1969.

[3] S. Sakata, "Finding a minimal set of linear recurring relations capable of generating a given finite 2D cyclic codes and their duals," *J. Symb. Comput.*, vol.5, pp.1321–1337, 1988.

[4] S. Sakata, "Extension of the BM algorithm to N dimensions," *Inform. & Comput.*, vol.84, pp.207–239, 1990.

[5] R. Kötter, "Fast GMD decoding of AG and RS codes," *IEEE Trans. on Inform. Theory*, vol.42, pp.721–737, 1995.

[6] U.K. Sorger, "A new RS codes decoding algorithm based on Newon's interpolation," *IEEE Trans. on Inform. Theory*, vol.39, pp.358–365, 1993.

[7] D.J. Taipale and M.J. Seo, "An efficient soft-decision RS decoding algorithm," *IEEE Trans. on Inform. Theory*, vol.40, pp.1130–1139, 1994.

[8] N. Kamiya, " On multisequence shift-register synthesis and GMD decoding of RS codes," *Finite Fields & Their Applic.*, vol.1, pp.440–457, 1995.

Queueing analysis with traffic models based on deterministic dynamical systems

Venkat Anantharam
EECS Department
University of California
Berkeley, CA 94720
ananth@vyasa.eecs.berkeley.edu

Abstract

Measurements of communication traffic from a wide variety of sources have suggested that correlations persist in such sources over fairly long time scales. This has called into question the use of traditional traffic models for performance analysis, and sparked an interest in studying novel kinds of traffic models. In particular, a suggestion has been made to build models based on expanding deterministic dynamical systems. Deterministic dynamical systems can exhibit chaotic behavior, which has many of the features of statistical behavior, and can indeed be studied rigorously using probabilistic techniques. We make some remarks regarding the analysis of queues driven by such traffic models.

1 Introduction

Statistical analyses of measurements of communication traffic from a wide variety of sources have suggested that correlations persist in such sources over fairly long time scales; for examples of such measurements, see [3], [16], and [21]. This has called into question the use of traditional traffic models for the performance analysis of communication networks, and has sparked considerable research on the use of "long-range dependent" traffic models for performance analysis; for examples see [1], [2], [14], [17], [19], and [20]. A feature of most of these works is the observation that the tail behavior of queues with long-range dependent inputs decays much slower than exponentially, either as a Weibull or according to a power law, depending on the model. The presence of a qualitative difference in queueing behavior with the traffic measured in [16] as compared to that predicted by conventional models is also supported by the experimental analysis of [11]. The persistence of long-range dependence in packet traffic even after it is regulated by simple flow control schemes

---

*Research supported by NSF NCR 9422513.

is demonstrated in [24]. A bibliography of work in this area as of the middle of 1996 is available in [26].

It should be pointed out that this area is quite active, and the jury is still out on the interpretation and significance of these results. There are at least two general grounds for caution. The first has to do with the very existence of "long-range dependence" (a nice discussion of a somewhat analogous earlier controversy in the field of geology is in [13]). For instance, the analysis of the Ethernet measurements in [16] is primarily done using the R/S statistic, variance-time plots, and the periodogram. The R/S statistic can be highly sensitive to trends in the observations; for instance, a special case of the main result of [4] is that the sum of an independent and identically distributed (i.i.d.) sequence of random variables with finite mean and variance and a deterministic trend that is vanishingly small and goes to zero asymptotically in time according to a power law, can result in the appearance of a nontrivial (not equal to 1/2) Hurst exponent when analyzed using the R/S statistic (for a precise formal theorem, see [4]). The behavior of variance-time plots for the Bellcore measurements of [16] under different kinds of nonstationarity is discussed in [8], and also in [23] (in the latter it is concluded that the measurements support the the hypothesis of long-range dependence fairly well). Certainly, the variance-time plots for the AUG89.MB measurements of [16] (plotted in Fig. 5 of [16]) would seem to support the existence of correlation up to a scale of about 100 seconds, under the hypothesis of stationarity, as a crude back of the envelope calculation of empirical variances for a stationary process model will show (see the appendix). This brings us to the second ground for caution, which is whether traditional traffic models with sufficiently large depth of correlation might already serve adequately for buffer dimensioning and network design purposes. Recent works that sound this note of caution include [12] and [25].

Whatever side the coin may fall on in this controversy, the measurements and statistical analyses of works such as [3], [16], and [21] are fascinating, and beg for some kind of physical explanation, whether it be due to trends, the nature of the underlying protocols, or the existence of "true" "long-range dependence". At least, such a physical understanding would aid in the construction of parsimonious and realistic traffic models, be they "traditional" or "long-range-dependent". One attempt at such a physical explanation is in [27], where it is argued via statistical analysis of the measurements in [16] that they are consistent with infinite variance of the load at the level of individual sources. It is proved in [22] that the superposition of ON/OFF sources whose ON-periods or OFF-periods have infinite variance results in an aggregate network traffic that is long-range dependent.

In this note our purpose is to make a couple of remarks in connection with another intriguing idea for arriving at physical descriptions of network communication traffic, suggested in [10].

2 Deterministic dynamical systems

The proposal in [10] is to model a communication traffic source (such as an Ethernet LAN, or a VBR video source) by means of a deterministic nonlinear transformation

$x_{n+1} = S(x_n)$, taking values in some state space X. The traffic source is modeled as having an activity level that depends on the current state.

Let us first remark that if X is allowed to be sufficient general, the restriction to deterministic S is not a big one. Indeed, the theory of deterministic chaos, see for instance [5] or [15], tells us that quite complicated statistical behavior can be expressed by such deterministic maps. For instance, consider the transformation on the unit interval $f : [0,1] \mapsto [0,1]$ given by

$$S(x) = \begin{cases} \frac{x}{p} & \text{if } 0 \leq x \leq p \\ 1 - \frac{x-p}{1-p} & \text{if } p \leq x \leq 1 \, . \end{cases}$$

Then, starting from Lebesgue almost any initial condition x_0, the empirical distribution of the sequence of iterates $\{S^n(x_0), n \geq 0\}$ will converge to Lebesgue measure on $[0,1]$ (in the weak topology of convergence of measures). Thus, there appears to be a stationary situation in which the state is distributed according to Lebesgue measure. Suppose the activity level of the source is

$$a(x) = \begin{cases} a & \text{if } 0 \leq x \leq p \\ 0 & \text{if } p \leq x \leq 1 \, . \end{cases}$$

If one were to start with the initial state distributed according to Lebesgue measure, the sequence of iterates $\{a(x_n), n \geq 0\}$ of the source values will be a sequence of i.i.d. $\{0, a\}$ valued Bernoulli random variables with probability of being a equal to p.

The point just made is that deterministic maps can be used to model stochastic processes also. Thus, by choosing X and the transformation S appropriately it should be possible to directly model the traffic generated by fairly complicated protocols and systems, in a directly tangible and physically meaningful way, while at the same time retaining the flexibility to incorporate traditional stochastic processes into the models. Another argument for the potential interest in such deterministic models is that there appears to be a significant amount of determinism in the structure of communication traffic - in Fig 1. of [10] the successive interarrival times from the Ethernet measurements of [16] is plotted, and visual inspection shows what appears to be considerable deterministic structure in this plot.

We now proceed to make a couple of remarks about the queueing behavior of queues driven by such traffic models.

3 Large deviations

We first record a special case of a result of [6], see Thm 3.9 (ii) of that paper. Let $\{a_n, n \geq 0\}$ be a stationary and ergodic sequence of nonnegative real valued random variables, interpreted as the sequence of arrivals into a single server queue, which can serve an amount c per unit time. Thus, the queue size evolves according to the equation

$$q_{n+1} = (q_n + a_n - c)^+ \, .$$

We assume that the stability condition $E[a_n] < c$ holds. From the result of [18], we know that the queue has a unique stationary distribution. Let q_∞ be a random variable with this distribution. We assume that for any θ, $0 < \theta < \infty$, the limit

$$a^*(\theta) \triangleq \frac{1}{n} \log E[e^{\theta \sum_{t=0}^{n-1} a_t}] \tag{1}$$

exists. Let us define

$$\theta^* \triangleq \sup\{\theta \ : \ a^*(\theta) < c\} .$$

Then the claim of [6] is that we have

$$\lim_{u \to \infty} \frac{-\log P(q_\infty \geq u)}{u} = \theta^* .$$

In this result, note that, since $\theta^* > 0$, the tail probability of the stationary queue size decreases exponentially.

We next note a special case of a result recorded in [7], see Theorem 6.4.4. and pg. 261 of that book. Given an $\mathbf{R}^K$- valued ψ-mixing sequence of random variables $\{Y_n, n \geq 0\}$, the sequence of empirical averages $\{S_n, n \geq 0\}$, where

$$S_0 \triangleq 0, \quad S_n \triangleq \frac{1}{n} \sum_{t=0}^{n-1} Y_t ,$$

obeys a large deviations principle with good convex rate function $\Lambda^*(\cdot)$, which is the Legendre transform of the function

$$\Lambda(\lambda) = \lim_{n \to \infty} \frac{1}{n} \log E[e^{n<\lambda, S_n>}] , \tag{2}$$

defined for $\lambda \in \mathbf{R}^K$, $\lambda = (\lambda_1, \ldots, \lambda_K)^T$. Here $< \lambda, S_n >$ denotes the inner product of vectors in $\mathbf{R}^K$. In particular, this result says that the limit in equation (2) exists.

4 Main remarks

Consider a deterministic dynamical system, given by a map $S : X \mapsto X$, where $(X, \mathcal{F}, \mu)$ is a measure space. A priori, the measure μ has nothing to do with the dynamics of S; it serves as a reference measure to permit talking about densities, and, as is customary, we will assume that S is measurable, and *nonsingular* with respect to μ, i.e. the induced measure $S_*(\mu)$ (which is also often written $\mu(S^{-1})$) is absolutely continuous with respect to μ.

We also assume given a partition of X into K measurable subsets $I_1, \ldots, I_K$, real numbers $a_1, \ldots, a_K$ (which need not be distinct), and define the function $a : X \mapsto \{a_1, \ldots, a_K\}$ by

$$a(x) = \sum_{k=1}^{K} a_k 1_{I_k}(x) ,$$

where $1_{I_k}(x)$ is the indicator function of I_k.

Suppose we start the dynamical system with a distribution that is absolutely continuous with respect to μ and can therefore be written as $f\mu$ for some density $f \in L^1(\mu)$ (i.e. $f \geq 0$ and $\int_X f(x)\mu(dx) = 1$). Then the next state will also have a distribution that is absolutely continuous with respect to μ (because S is nonsingular with respect to μ), so there is an operator P on $L^1(\mu)$ such that $P(f)$ is the density of this next state with respect to μ. P is called the *Frobenius - Perron operator* of the transformation.

Under a wide range of conditions it is known that, for arbitrary initial density f, the sequence of iterates $\{P^n(f), n \geq 0\}$ will converge to a limit f^* such that $P(f^*) = f^*$. For details regarding the sense in which convergence takes place, and situations in which it is know to take place, see [5] and [15]. There is a vast literature this topice, with conditions known that cover several nonlinear maps with quite nontrivial behavior. For maps of the unit interval $[0,1]$, for instance, a commonly cited example of such a result is recorded in Theorem 6.2.1 of [15]. Thus, the state of the dynamical system appears to converge, in some appropriate sense, to a random state with distribution $f^*\mu$. If one starts the dynamical system with this distribution, then the sequence of arrivals $\{a_n, n \geq 0\}$ will be a stationary ergodic process. It is certainly the case that under a wide range of conditions, this process will be ψ-mixing; indeed, the underlying dynamical system itself can be shown to be strong mixing in many cases; for instance, *exactness* of the transformation, which is often quite easy to see, implies mixing.

Combining these observations with the results stated in the preceding section leads to our first remark : If one constructs such an expanding discrete dynamical system model, and an arrival process derived from such a model, then, in the (non-pathological) situation where a stationary distribution exists for the dynamical system, one should typically expect exponential decay of the tails of the stationary queue size in queues driven by the arrivals. It suffices to apply the second result of section 3 to the sequence of $\mathbf{R}^K$ valued random variables

$$\{(1_{I_1}(x_n), \ldots, 1_{I_K}(x_n)),\ n \geq 0\}\ ,$$

to conclude the existence of the limit in equation (1).

We next observe that positive recurrent discrete time finite state Markov chain arrivals are quite easily constructed along the lines of the general scheme we are discussing. To get a stationary sequence $\{a_n, n \geq 0\}$ taking values in $\{a_1, \ldots, a_K\}$ with stationary distribution $(\pi_k, 1 \leq k \leq K)$ and transition probability matrix $(p_{ij}, 1 \leq i, j \leq K)$, take X to be the unit interval $[0,1]$, $\mathcal{F}$ to be the Borel σ-field, and μ to be Lebesgue measure. Take $I_1, \ldots, I_K$ to be intervals of lengths $\pi_1, \ldots, \pi_K$ respectively, and further partition each I_k into intervals I_{kl}, $1 \leq l \leq K$, of lengths $\pi_k p_{kl}$ respectively. Consider the deterministic transformation

$$S : [0,1] \mapsto [0,1]$$

which is *piecewise linear* and maps the interval I_{kl} onto the interval I_l (so that it has slope $\frac{\pi_l}{\pi_k p_{kl}}$ on this interval. Except for the quibble that a state of the Markov chain may lead to another state of the chain with probability 1, which situation can also be easily handled, this is an example of a piecewise linear (expanding) Markov transformation, in the sense of Chapter 9 of [5], because the slopes are all strictly bigger than 1; it is easily seen to leave Lebesgue measure invariant (so the invariant density is 1). The sequence of arrivals $\{a_n, n \geq 0\}$, when we start the dynamical system with Lebesgue measure, will be a Markov chain with initial distribution $(\pi_k, 1 \leq k \leq K)$ and transition probability matrix $(p_{ij}, 1 \leq i, j \leq K)$.

Our second remark is the following : Suppose we are given any piecewise expanding transformation of the interval (for the definition, see page 85 of [5]). These are among the most studied transformations in the literature, and for maps of the interval an expanding condition is required for most results regarding the existence of a stationary distribution. [1] Then one can approximate the transformation by a piecewise linear (expanding) Markov transformation. The arrival process corresponding to this approximation, in stationarity, is a function of a stationary finite state Markov chain. The limit of equation (1) can now be *explicitly* written down in terms of the slopes of the approximation to the transformation. This is a basic result that goes back to the seminal works of Donsker and Varadhan for the large deviations of the empirical distribution of Markov processes; for a convenient reference for such results, see [9]. Thus, according to the result of [6], the tail probabilities of a single server queue driven by such an arrival process will decay exponentially (assuming stability), and the rate of decay of the tail can be explicitly computed in terms of the parameters of the approximating Markov transformation.

Note that our second remark can be generalized to transformations of spaces other than the interval - all one needs is that it should be possible to construct a finite state symbolic dynamical scheme that approximates the given transformation.

5 Conclusion

Deterministic dynamical systems offer a promising (or at least suggestive !) modeling paradigm for the communication traffic processes generated by systems executing complex protocols, which might also be composed of several interacting components. It is important to realize that one has *not* lost modeling flexibility by restricting to deterministic transformations, because stochastic processes can also be modeled by this paradigm. What one appears to gain is a direct physical approach to building models based on an explicit description of the underlying system dynamics. We have pointed out that when we consider a queue fed by arrivals derived from such a system, we should typically expect to see exponentially decaying tails for the stationary queue size. Further, by chosing suitable approximations for the transformation describing the system, we can write explicit formulas for the rate of decay of the tail probabilities of the stationary queue size.

---

[1] The models proposed in [10] are also of this type, although the focus there is on conjectures regarding the situation corresponding to "intermittency" in physics, where the expanding condition fails to hold at one point.

6 Appendix

Let $X_1, \ldots, X_N$ be a sequence of conditionally independent Poisson random samples, conditioned on means $\mu_1, \ldots, \mu_N$ respectively. Assume $\mu_n = \lambda + \gamma_n$, $1 \leq n \leq N$. This is a simple model for a sequence of packet or byte counts with fluctuating mean. The samples are aggregated at level m to get

$$X_j^{(m)} = X_{(j-1)m+1} + \ldots + X_{jm} \; , \quad 1 \leq j \leq M \; ,$$

where $1 \leq m \leq N$, m divides N, and M denotes N/m. The variance time plot is the plot of the logarithm of the empirical variance of the m-aggregated sequence $X_1^{(m)}, \ldots, X_M^{(m)}$ against $\log m$.

Write Γ_n for $\sum_{t=1}^{n} \gamma_t$. The conditional expectation of the empirical variance of the m-aggregated sequence is seen to be

$$\frac{\lambda + \Gamma_N/N}{m} - \frac{\lambda + \Gamma_N/N}{N} + \frac{1}{Nm}\sum_{j=1}^{M}(\Gamma_{jm} - \Gamma_{(j-1)m})^2 - \frac{{\Gamma_N}^2}{N^2} \; .$$

Thus, loosely speaking, if $\Gamma_N = o(N)$, and, for most j, $(\Gamma_{jm} - \Gamma_{(j-1)m})^2 \sim m^{2H}$ over the range $m_0 \leq m \leq \frac{N}{m_0}$, then the variance time plot will exhibit a slope of $1 - 2H$ over this range of m. On the basis of this crude calculation the variance time plot of Fig. 5(b) of [16] could be taken to indicate correlations on the scale of about 100 seconds for the AUG89.MB measurements described there, but do not necessarily indicate correlations on longer time scales than this.

References

[1] Anantharam, V. (1995). On the Sojourn Time of Sessions at an ATM Buffer with Long-Range Dependent Input Traffic. *Proceedings of the 34rd IEEE Conference on Decision and Control*, New Orleans, Louisiana, December 13 -15, 1995, Volume 1, pp. 859 -864.

[2] Anantharam, V. (1996). "Networks of Queues with Long-Range Dependent Traffic Streams", *Stochastic Networks: Stability and Rare Events*, edited by Paul Glasserman, Karl Sigman, and David Yao, Lecture Notes in Statistics, Vol. 117, Springer Verlag, pp. 237 - 256.

[3] Beran, J., Sherman, R., Taqqu, M.S., and Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, Feb.-April 1995, vol.43, Nos. 2-4, pt.3, pp. 1566 -1579.

[4] Bhattacharya, R. N., Gupta, V. K., and Waymire, E. (1983). The Hurst Effect under Trends. *Journal of Applied Probability*, Vol. 20, pp. 649 -662.

[5] Boyarsky, A. and Góra, P. (1997). *Laws of Chaos : Invariant Measures and Dynamical Systems in One Dimension.* Birkhäuser, Boston.

[6] Chang, C. S. (1994). Stability, Queue Length, and Delay of Deterministic and Stochastic Networks. *IEEE Transactions on Automatic Control*, Vol. 39, No. 5, pp. 913 -931.

[7] Dembo, A. and Zeitouni, O. (1993). *Large Deviation Techniques and Applications.* Jones and Bartlett Publishers, Boston.

[8] Duffield, N.G., Lewis, J.T., O'Connell, N., Russell, R., and Toomey, F. (1994). Statistical issues raised by the Bellcore data. *IEE Eleventh UK Teletraffic Symposium. Performance Engineering in Telecommunications Networks*, Cambridge, UK, Vol 1 pp. 1-8.

[9] Ellis, R. S. (1988). Large Deviation for the Empirical Measure of a Markov Chain with an Application to the Multivariate Empirical Measure. *Annals of Probability*, Vol. 16, pp. 1496 -1508.

[10] Erramilli, A. and Singh, R. P. (1995). An application of deterministic chaotic maps to model packet traffic. *Queueing Systems : Theory and Applications*, Vol 20, pp. 171 -206.

[11] Erramilli, A., Narayan, O., and Willinger, W. (1996). Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, April 1996, vol.4, no.2, pp. 209-23.

[12] Heyman, D.P. and Lakshman, T.V. (1996). What are the implications of long-range dependence for VBR-video traffic engineering?. *IEEE/ACM Transactions on Networking*, June 1996, vol.4, no.3, pp. 301 -317.

[13] Klemes, V. (1974). The Hurst phenomenon : a puzzle ? *Water Resources Research*, Vol. 10, pp. 675 -688.

[14] Konstantopoulos, T. and Lin, S-J. (1996). Fractional Brownian Motion and Lévy Motions as Limits of Stochastic Traffic Models. *Proceedings of the 34th Annual Allerton Conference on Communications, Control, and Computing*, Urbana, Illinois, pp. 913 -922.

[15] Lasota, A. and Mackey, M. C. (1994). *Chaos, Fractals, and Noise : Stochastic Aspects of Dynamics.* Springer Verlag, New York.

[16] Leland, W. E., Taqqu, M. S., Willinger, W. and Wilson, D. V. (1994). On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, pp. 1-15.

[17] Likhanov, N., Tsybakov, B. and Georganas, N. D. (1995). Analysis of an ATM Buffer with Self-Similar ("Fractal") Input Traffic. *Proceedings of the 14th Annual IEEE Infocom*, pp. 985-992.

[18] Loynes, R. M. (1962). The Stability of Queues with Non-Independent Interarrival and Service Times. *Proceedings of the Cambridge Philosophical Society*, Vol. 58, pp. 497-520.

[19] Norros, I. (1994). A storage model with self-similar input. *Queueing systems : Theory and applications*, Vol. 16, pp. 387 -396.

[20] Parulekar, M. and Makowski, A. M. (1996). Tail Probabilities for a Multiplexer with Self-similar Traffic. *Proceedings of the IEEE INFOCOM 1996*, pp. 1452 -1459.

[21] Paxson, V. and Floyd, S. (1995). Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp. 226-244.

[22] Taqqu, M.S., Willinger, W., and Sherman, R. (1997). Proof of a fundamental result in self-similar traffic modeling. *Computer Communication Review*, April 1997, vol.27, no.2, pp. 5-23.

[23] Teverovsky, V. and Taqqu, M. (1995). Testing for Long-Range-Dependence in the Presence of Shifting Means or a Slowly Declining Trend, using a Variance-type Estimator. *Journal of Time Series Analysis*, Vol. 18, No. 3, pp. 270 -304.

[24] Vamvakos, S. and Anantharam, V. (1998). On the Departure Process of a Leaky Bucket System with Long-Range Dependent Input Traffic. *Queueing Systems : Theory and Applications*, 1998. to appear. Available as *UCB/ERL Memorandum*, No. M96/51 and *UCB/ERL Memorandum*, No. M96/70, Electronics Research Laboratory, University of California, Berkeley.

[25] Venkateshwara Rao, B., Krishnan, K.R., and Heyman, D.P. (1996). Performance of finite-buffer queues under traffic with long-range dependence. *Proceeding of the IEEE GLOBECOM 1996*, London, UK, vol.1., pp. 607 -611.

[26] Willinger, W., Taqqu, M.S., and Erramilli, A. (1996). A Bibliographical Guide to Self-similar Traffic and Performance Modeling for Modern High-speed Networks. In *Stochastic Networks : Theory and Applications*, edited by F. Kelly, S. Zachary, and I. Ziedins, Oxford University Press, Oxford, pp. 339 -366.

[27] Willinger, W., Taqqu, M.S., Sherman, R., and Wilson, D.V. (1997). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking,* Feb. 1997, vol.5, no.1, pp. 71 -86.

Analytic Information Theory in Service of Queueing with Aggregated exponential On/Off Arrivals

Philippe JACQUET

September 22, 1997

INRIA-Rocquencourt
Domaine de Voluceau
BP 105 78153 Le Chesnay
FRANCE

Abstract

We present a simple and realistic queueing model in which a time arival process is made of an aggregation of several independent on/off sources. We show that such a process generates asymptotic self-similaritites or long term correlations with the time. It also shows that this arrival process when applied to a single server generates queue length with at least polynomial tail.

1 Introduction

Our primary aim is to illustrate the use of some tools borrowed from Analytic Information Theory to the specific queueing model involving aggregations of independent "on/off" input sources. This model finds application in networking.

We coin here the term *Analytical Information Theory* to describe problems of information theory that are solved by analytical methods, that is, those in which complex analysis plays a primary role [7, 10]. Look at [11] for a survey and at [1] for a detailled description of the tools and proofs used in the present paper.

Let us consider a large multiple-access network like an Ethernet link or an ATM switch. By "large" we mean a large population of users and by "multiple-access" we mean that a single resource of communication has to be shared. We assume that users are independent from each other, which is a reasonnable assumption. We also assume that every user behaves like an "on/off" input source.

"On/off" input sources are very useful for real networking modelisation. User activity is characterized by *wake* periods interleaved with *sleep* periods. During a wake period, the user has intense activity, *i.e.* creates messages with constant peak rate, λ, (*e.g.* for files retrieval, network browsing, *etc*). When the user goes for a coffee break, or works on local device, or simply when he goes home, he is in sleep period. During a sleep period, the user has very low activity, to simplify we will assume no activity at all.

To simplify we also suppose that the multiple access communication resource is a queueing with a single server, and messages plays the role of customers stored in this queueing waiting for service.

If the sleep/wake periods follow very similar distributions from one user to another, then the aggregated input process would benefit of renewal properties which will make it very stable. When the number of users tends to infinity such input process would asymptotically look like a global continuous Poisson process with well known stability conditions and distribution parameters.

Here, we are interested in sleep/wake cycles that follow very different distributions from one user to another. This very likely happens in a real netwok. For example it seems that people which use telephone less often than other, in general have longer calls, (Grand' Ma Christmas Call syndrom). In other words users with lower overall activity have in general much longer wake periods than users with higher activity. The fact that such users have finally lower overall activity is recovered by the fact that their sleep periods are proportionnaly much more longer.

If the discrepancies between user sleep/wake period distributions fit a polynomial bound, described in the paper, we will prove *via* analytic methods the two following main results:

1. The overall arrival process is asymptotically self-similar.
2. The queue distribution has at least polynomial tail.

The first may provide a (very superficial) surprise, since the aggregation of independent sources with renewal properties should not produce long term correlations with arbitrarily strong amplitude. However this will occurs simply because the scaling discrepancies between the sources are too important as it is usual to find in *fractal* analysis. The second result confirms statistics about buffer occupancy in network under real traffic conditions (Bellcore statistical data [6]).

The impact of self-similarities in a network should not be neglected in practical implementations. Clearly basic uniform Poisson models lead to queue distributions with exponential tails. When this tail is known, engineer can specify the minimum buffer size which fits a maximum overflow probability requirement. But if the tail turns out to be polynomial, then the designer would need to rise the previous minimum buffer size estimate to some exponential order to fit the same overflow probability requirement.

In the following we will assume that in each single on/off source the wake and sleep period durations are exponentially distributed with respective parameter Poisson-wake and Poisson-sleep. We then consider a set of on/off sources such that the sequence of the ratio of Poisson-wake and Poisson-sleep decays like polynomials of degree β. We will show that under some conditions that the arrival process obtained by the aggregation of the on/off sources is β-similar (section 2) and the queue length has a polynomial tail of degree at least β (section 3).

2 Self similarities in aggregated on/off sources

2.1 Definition of the on/off source

As we said in introduction, an on/off process is a Poisson arrival process with a time varying rate. The Poisson rate oscillates between0 and λ. The transitions times between these two values are exponentially distributed with respective values ν_1 and ν_0. Quantity ν_1 is the transition rate from state "up" to state "down" and quantity ν_0 is the transition rate from state "down" to state "up". In other words, ν_1 is the Poisson-wake parameters and ν_0 is the Poisson-sleep parameter. See [2, 1] for a detailed description.

2.2 Time covariance in a single on/off source

Let I be a time interval of size θ and for x non-negative let $I+x$ the same interval translated of $\theta + x$. We define $A(I)$ as the number of arrivals that occurs during interval I. We define the quantity $\mathrm{Cov}(\theta, x)$ as the covariance of $(A(I), A(I+x))$, *i.e.* $\mathrm{Cov}(\theta, x) = \mathrm{E}(A(I) \times A(I+x)) - \mathrm{E}(A(I))\mathrm{E}(A(I+x))$. Our aim is to give a closed formula for $\mathrm{Cov}(\theta, x)$.

To this end we introduce the following matrix T that we call the *transition* matrix:

$$\mathbf{T} = \begin{bmatrix} -\nu_1 & \nu_0 \\ \nu_1 & -\nu_0 \end{bmatrix} \tag{1}$$

The eigenvalues of $\mathbf{T}$ are 0 and $-(\nu_1 + \nu_0)$.

Let us suppose that at time 0, we start the process with a state randomly distributed between "on"s and "off"s according to a given probability vector $\mathbf{u}$. Via an easy Markov model we compute the probabilistic repartition between state "up"s and state "down"s of the arrival process at a further time x. The probability vector is exactly $\exp(x\mathbf{T})\mathbf{u}$. Notice that the eigenvalues of $\exp(x\mathbf{T})$ are the exponential of the eigenvalues of $\mathbf{T}$ time x, *i.e.* 1 and $\exp(-(\nu_1 + \nu_0)x)$.

Going further, we establish that the expected proportion of "up"s periods and "down" periods intercepted by the interval $[0, \theta]$ will be $1/\theta \int_0^\theta \exp(x\mathbf{T})\mathbf{u}dx$. Notice that the eigenvalues of $\int_0^\theta \exp(x\mathbf{T})\mathbf{u}dx$ are respectively θ and $(1 - \exp(-(\nu_1 + \nu_0)\theta))/(\nu_1 + \nu_0)$. If we define ℓ as the covector $[\lambda, 0]$, then the average arrival on interval $[0, \theta]$ will be $\langle \ell | \int_0^\theta \exp(x\mathbf{T})\mathbf{u}dx \rangle$.

Theorem 1 *We have the following explicit formula:*

$$\begin{aligned} Cov(\theta, x) \quad = \quad & [\lambda(1 - \exp(-(\nu_1 + \nu_0)\theta))]^2 \times \\ & \times \exp(-(\nu_1 + \nu_0)x)\nu_1\nu_0(\nu_1 + \nu_0)^{-4}. \end{aligned} \tag{2}$$

Proof: If we fix θ, then the function $\mathrm{Cov}(\theta, x)$ is a linear function of $\exp(x\mathbf{T})$ and therefore a linear combination of the two eigenvalues 1 and $\exp(-(\nu_1 + \nu_0)x)$. Identification of the coefficients easily comes. ∎

2.3 Aggregations of on/off arrival sources

In the following we consider a general system where the arrival process is a superposition of several independent on/off sources. We consider a denumerable set of on/off sources indexed from 1 to ∞. See figure 1. Let I be an interval, we call $A_j(I)$ the number of arrivals coming from the jth source during interval I. Let $X(I)$ denotes the total arrival numbers queued during interval I we obviously have $X(I) = \sum_{j=1}^{\infty} A_j(I)$.

In the sequel we will show that such an aggregation can lead to an asymptotic self-similar arrival process. An arrival process is β self-similar if for any fixed interval I, quantity $\mathrm{Cov}(X(I), X(I+x))$ is of order $x^{-\beta}$, when x tends to infinity. It is said to be at most β self-similar if $\mathrm{Cov}(X(I), X(I+x))$ is of magnitude order at most $x^{-\beta}$.

To simplify the computations througout this section we assume that:

- for every i, the ith on/off source has peak rate λ_i;
- the λ_i are all identical and equal to a given $\lambda > 0$;

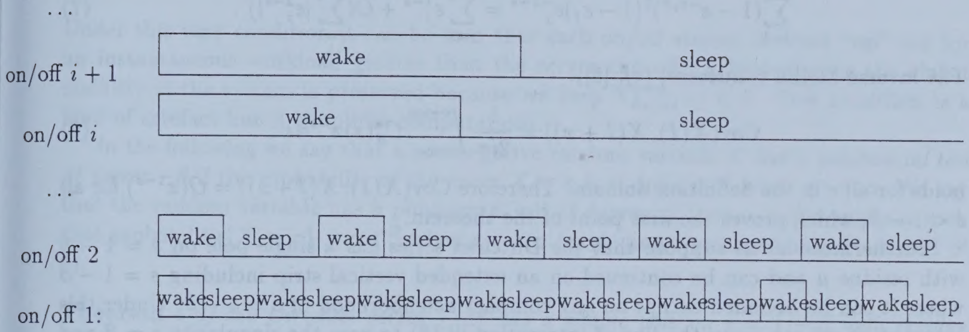

...
on/off $i+1$
wake
sleep
on/off i
wake
sleep
...
on/off 2
wake
sleep
wake
sleep
wake
sleep
wake
sleep
wake
sleep
on/off 1:
wake
sleep
wake
sleep
wake
sleep
wake
sleep
wake
sleep
wake
sleep
wake
sleep
wake
sleep

Figure 1: Aggregation of a denumerable set of independent on/off sources

- for every on/off source we have $\nu_0 = \varepsilon_i^2$ and $\nu_1 = \varepsilon_i - \varepsilon_i^2$, for some sequence $\varepsilon_i > 0$.

Consequently, the average arrival per unit time is $\lambda \times \sum_{j=1}^{\infty} \varepsilon_j$.

In the sequel we suppose that the series in $\sum \varepsilon_i$ is convergent. We also assume that the Dirichlet series $\sum_{j=1}^{\infty} \varepsilon_j^s$ is absolutely convergent for all complex numbers s with real part strictly less than $1-\beta$. For example $\varepsilon_j = j^{1/(1-\beta)}$. Notice that in this case the Dirichlet series can be identified with the Riemann ζ function.

Theorem 2 *If the serie $\sum \varepsilon_j^{1-\beta}$ converges, then the aggregated process is at most β self-similar. Furtheremore, if the Dirichlet series has a simple pole on $s = 1-\beta$ with residue μ then the aggregated process is β self-similar and the coefficients are exactly determined.*

Proof: Let us choose an interval I of size θ. Using the independence of the A_i we obtain:

$$\mathrm{Cov}(X(I), X(I+x)) = \sum_{j=1}^{\infty} \mathrm{Cov}(A_j(I), A_j(I+x)) \tag{3}$$

Referring to Theorem 1 we obtain:

$$\mathrm{Cov}(X(I), X(I+x)) = \lambda^2 \sum_{j=1}^{\infty} (1-e^{-\varepsilon_j\theta})^2 (1-\varepsilon_j)\varepsilon_j^{-1} \exp(-\varepsilon_j x) \tag{4}$$

We identify in the above a harmonic sum [5]. The asymptotics of such a sum is classically obtained via the analysis of the definition domain and the singularities of its Mellin transform $C_\theta^*(s)$, that is, defined for appropriate complex numbers s by:

$$C_\theta^*(s) = \int_0^{\infty} \mathrm{Cov}(X(I), X(I+x)) x^{s-1} \mathrm{d}x \ . \tag{5}$$

From 4 we thus obtain (*cf.* [5]):

$$C_\theta^*(s) = \lambda^2 \Gamma(s) \sum_{j=1}^{\infty} (1-e^{-\varepsilon_j\theta})^2 (1-\varepsilon_j)\varepsilon_j^{-1-s} \tag{6}$$

where $\Gamma(s)$ is the Mellin transform of function e^{-x}.

The expression of $C^*(s)$ converges for all s such that $\Gamma(s)$ and the series ε_j^{1-s} converges, *i.e.* for $0 < \Re(s) \leq 1 - \beta$. Indeed:

$$\sum_{j=1}^{\infty}(1 - e^{-\varepsilon_j\theta})^2(1 - \varepsilon_j)\varepsilon_j^{-1-s} = \sum_{j=1}^{\infty} \varepsilon_j^{1-s} + O(\sum_{j=1}^{\infty} |\varepsilon_j^{2-s}|) \ . \tag{7}$$

The inverse Mellin transform, (*cf.* [5]):

$$\mathrm{Cov}(X(I), X(I+x)) = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty} C_\theta^*(s)x^{-s}\mathrm{d}s \ , \tag{8}$$

holds for all c in the definition domain. Therefore $\mathrm{Cov}(X(I), X(I+x)) = O(x^{1-c})$ for all $c < 1-\beta$, which proves the first point of the theorem.

Furthermore let us suppose that the Dirichlet series has a single pole on $s = 1-\beta$ with residue μ and can be continued on an extended vertical strip including $s = 1-\beta$ where it can be bounded except on any compact neighborhood of $s = 1-\beta$. Under this assumption we can move the line of integration in (8) to pass the singularity $s = \beta$ and to stand ε further on the right. The residue theorem gives:

$$\mathrm{Cov}(X(I), X(I+x)) = \mu x^{-\beta} + \int_{\beta+\varepsilon-i\infty}^{\beta+\varepsilon+i\infty} C_\theta^*(s)x^{-s}\mathrm{d}s \tag{9}$$

The last integration on the left hand side is $O(x^{-\beta-\varepsilon})$ via obvious majoration under the sign "$\int$". ∎

Remark: One can determine oscillatory coefficients in $\log x$ just by simple tuning of system parameters. Such details are left out of the scope of the present paper.

3 Queueing with on/off input sources

In the following, the network is modelled via a queue with a single server which receives input from a set of on/off sources. The service time is exponential with mean 1 time unit. Our aim is to find the steady state distribution of the queue length and in particular to find asymptotics of the probability p_n that the queue length exceeds n customers, when $n \to \infty$.

3.1 Queueing with a single on/off source

We refer to [2] to state the following theorem:

Theorem 3 *The queue length generating function with the above on/off input source satisfies:*

$$q(z) = \frac{(1-\lambda\varepsilon z_1)z - (1-\lambda\varepsilon)z_1}{z - z_1} \tag{10}$$

with

$$z_1 = 1/2\frac{1+\lambda+\varepsilon+\sqrt{1-2\lambda+2\varepsilon+\lambda^2+2\lambda\varepsilon+\varepsilon^2-4\lambda\varepsilon^2}}{\lambda+\lambda\varepsilon^2} \ . \tag{11}$$

Proof: This a straightforward adaptation of [2]. ∎

Corollary 1 *When $n \to \infty$, quantity p_n exponentially decreases and $p_n = \lambda\varepsilon z_1^{1-n}$.*

3.2 Queueing with under an infinite number of on/off sources

We adopt the same hypotheses as in Section 2.3, and we add a new condition:

- The quantity λ is strictly greater than 1.

Under this very condition it can be seen that each on/off stream in state "up" ask for an instantaneous workload greater than the server can offer. Nevertheless the global stability of the system is preserved because we keep $\lambda \sum_{j=1}^{\infty} \varepsilon_j < 1$. This condition is a kind of artefact but it simplifies computations.

In the following we say that a non-negative random variable X has a *polynomial tail of degree* $-\beta$ if the probability of the event $X > x$ is of order $x^{-\beta}$ when $x \to \infty$. We say that the random variable has a polynomial tail of degree at least $-\beta$ if it exists α such that probability$\{X > x\} > \alpha x^{-\beta}$ when $x \to \infty$ (excluding a compact neighbourhood of 0).

Theorem 4 *If the Dirichlet series* $\sum_i \varepsilon_i^s$ *has a simple pole on* $s = 1 - \beta$ *with residue* μ, *then the queue length has a polynomial tail of degree at least* $-\beta$ *and coefficient for a lower bound can be exactly determined.*

Proof: To simplify, we assume the same hypotheses as in the proof of Theorem 2: the Dirichlet series can be continued to a vertical strip $1 - \beta < \Re(s) < 1 - \beta + \varepsilon$.

Instead of directly looking on the queue length in the single server mode we consider a modified system called "multi-server" system. In the multi-server system we assume that every on/off stream has a separate queue and a distinct server. It is clear that the sum of the queue length in the multi-server system is always smaller than the queue length of the single server system. Therefore if the multi-server system shows a polynomial tail of degree $-\beta$ then the single server system will show a polynomial tail of degree at least $-\beta$.

Our aim is to show that the multi-server model shows a polynomial tail of degree $-\beta$. Let $g(z)$ the probability generating function of the sum of the queue lengths of the multi-server mode. We have the obvious identity: $g(z) = \prod_{i=1}^{n} q_i(z)$, where $q_i(z)$ is the probability generating function of the queue size of the server attached to the on/off stream number i.

Generating functions $q_i(z)$ are computed according to Theorem 3. Let z_i be the equivalent of z_1 in theorem 3 for on/off source i. We have formally:

$$g(z) = \prod_{i=1}^{\infty} \frac{(1 - \lambda\varepsilon_i z_i)z - (1 - \lambda\varepsilon)z_i}{z - z_i} \tag{12}$$

Quantity $g(z)$ is defined as long as z does not belong to the set of singularities z_i. Under this restriction, the expression $g(z)$ converges because the series in $z_i - 1$ converges. Indeed, $z_n = 1 + \varepsilon_n/(\lambda - 1) + O(\varepsilon_i^2)$ when $n \to \infty$, and the series in ε_i converges. In passing $g(z) = a + O(1/z)$ with $a = \prod_{i=1}^{\infty}(1 - \lambda\varepsilon_i z_i)$, when $|z| \to \infty$.

In the following we will establish that $g(1-x) = 1 + \alpha x^{\beta} + O(x^{\beta+\varepsilon})$ when x converges to 0 with the condition that $|\arg(z)| < \pi - \theta$ for some α and $\theta < \pi/2$. We define $\ell(z) = \log(g(1-x)) - \log(a)$. We have

$$\ell(x) = \sum_{i=1}^{\infty} \left[\log(1 + \frac{1 - \lambda\varepsilon_i z_i}{z_i - 1}x) - \log(1 + \frac{x}{z_i - 1}) \right] - \log(a) \; . \tag{13}$$

Since $\ell(x) = -\log(a) + o(1)$ when $x \to 0$ and that $\ell(x) = O(1/x)$ when $x \to \infty$, the Mellin transform $\ell^*(s)$ of $\ell(x)$ is defined on the strip $0 < \Re(s) < 1$ and has expression (*cf.* [5]):

$$\sum_{i=1}^{\infty} \left[(\frac{1-\lambda\varepsilon_i z_i}{z_i - 1})^{-s} - (\frac{1}{z_i - 1})^{-s} \right] \frac{-\pi}{s \sin \pi s} . \tag{14}$$

We identify in $(s \sin \pi s)^{-1}(-\pi)$ the Mellin transform of function $\log(1+x)$.

To get asymptotics of $\ell(x)$ when $x \to 0$ we identify the singularities of $\ell^*(s)$ located on the left of the definition domain, i.e for negative $\Re(s)$ in order to use the residu theorem in the reverse Mellin transform.

The first encountered pole is at $s = 0$ since $(s \sin \pi s)^{-1}\pi$ has double root at $s = 0$. This pole is finally of degree 1 since the factor in front of it has also root at $s = 0$, its residue is $\sum_{i=1}^{\infty} \log(1 - \lambda\varepsilon_i z_i) = \log(a)$. There is a less obvious pole on $s = -\beta$. Indeed:

$$\sum_{i=1}^{\infty} \left[(\frac{1-\lambda\varepsilon_i z_i}{z_i - 1})^{-s} - (\frac{1}{z_i - 1})^{-s} \right] = \sum_{i=1}^{\infty} s\varepsilon_i^{1+s}(1-\lambda)^{-s}(1 + O(\varepsilon_i)) \tag{15}$$

since $z_i = 1 + \varepsilon_i/(\lambda - 1) + O(\varepsilon_i^2)$. Therefore the pole at $s = -\beta$ has residu $\mu\beta(1-\lambda)^{\beta}$.

Applying the residu theorem in the reverse Mellin transform by moving the line of integration from c to $-\beta - \varepsilon$:

$$\begin{aligned} \ell(x) &= \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} \ell^*(s) x^{-s} ds \\ &= -\log(a) - \mu\beta(1-\lambda)^{\beta}\pi(\beta \sin \pi\beta)^{-1} x^{\beta} + \\ &\quad + \frac{1}{2i\pi} \int_{-\beta-\varepsilon-i\infty}^{-\beta-\varepsilon+i\infty} \ell^*(s) x^{-s} ds \end{aligned}$$

A trivial majorization under the above integral, translates immediately into the correct asymptotics of $g(z)$ around $z = 1$: $g(1-x) = 1 - \alpha x^{\beta} + O(x^{\beta+\varepsilon})$ with $\alpha = a\mu\beta(1-\lambda)^{\beta}\pi(\beta \sin \pi\beta)^{-1}$. Of course, this asymptotic is valid under the condition that x varies in an open set such that $|\arg(x)| < \pi - \theta$ for a θ arbitrarily chosen $\theta < \pi/2$.

The latter conditions suffice to translate asymptotics of $g(z)$ into asymptotics over its coefficients. Let b_n be the coefficients of $g(z)$: $\sum_{n=0}^{\infty} b_n z^n = g(z)$. Since $g(1-x) = 1 - \alpha x^{\beta} + O(x^{\beta+\varepsilon})$. We immediately deduce from the singularity analysis of [4] that

$$b_n = \frac{\alpha}{\Gamma(-\beta)} n^{-1-\beta} + O(n^{-1-\beta-\varepsilon}) . \tag{16}$$

The cumulative coefficients $p_n = b_n + b_{n+1} + \cdots$ observe similar asymptotics:

$$p_n = \frac{\alpha}{\Gamma(1-\beta)} n^{-\beta} + O(n^{-\beta-\varepsilon}) . \tag{17}$$

The above expression proves the theorem. ■

Remark: We can make a similar analysis when the singularity set is a little more complicated. In principle, it is possible to obtain tail distributions with oscillating terms. We can also treat more general cases. The artefact condition $\lambda > 1$ can also be eliminated. But the most promising workitem left is the estimation of a suitable *upper* bound for the queue length tail distribution, since the present work is limited to lower bounds.

4 Simulation Results

We analyse an aggregation of an infinite sequence of independent on/off sources. In fact we use an aggregation of 60 on/off sources. The parameters of the on/off input with index i are: $\lambda_i = 2.0$, $\varepsilon_i = \alpha i^{1/(\beta-1)}$; we choose $\beta = 0.5$ and β is tuned in order to keep the average input at 0.6. Figure 2 displays the sequence of on/off parameters on a logarithmic scale.

Figure 3, displays the distribution of queue size. The results are obtained by sampling 20,000 windows of 10 time unit size. If we select a logarithmic scale for horizontal axis, then the curve $\log(p_n)$ versus $\log n$ looks linear as a kind of illustration of a polynomial tail distribution.

5 Conclusion

We have presented a simple realistic model of arrivals based on the on/offs aggregation. We show that appropriate tuning of parameters produces time self-similariries and polynomial tail distributions buffer size. The results are nevertheless partial. Future work can progress in two different directions. The first direction is to find less restrictive conditions on parameter tuning. The second direction is to find an exact estimate of the asymptotic tail distribution instead of a lower bound.

References

[1] P. Jacquet, Queueing with aggregated up/down arrivals, INRIA RR-2663, 1995.

[2] M. Neuts, *Matrix-geometric solutions in stochastic models : an algorithmic approach*, the John Hopkins University press, 1981.

[3] N. Likhanov, B. Tsybakov, N. Georganas, Analysis of an ATM buffer with self-similar ("fractal") input, submitted, 1994.

[4] P. Flajolet, A. Odlyzko, Singularity analysis of generating functions, in *SIAM J. Disc. Math.*, Vol 3, No 2, pp. 216-240, 1990.

[5] P. Flajolet, X. Gourdon, P. Dumas, Mellin transform and asymptotics: Harmonic sums, in *Theoretical Computer Science*, Vol 144, No 1-2, pp. 3-58, 1995.

[6] W. Willinger, M. Taqqu, W. Leland, D. Wilson, Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurement, in *Statistical Science*, Vol 10, No 1, pp. 67-85, 1995.

[7] P. Jacquet, W. Szpankowski, Toward analytical information theory: entropy computations, submitted, 1997.

[8] P. Glynn, W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, *J. Appl. Prob.*, to appear.

[9] N. Duffield, N. O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. DIAS Technical Report No DIAS-STP-93-30, 1993.

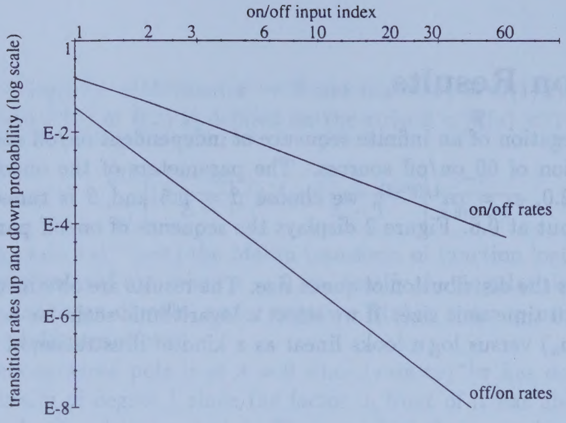


Figure 2: Transition rates versus on/off index. Transition rates from on state to off state: ν_i (on/off rates); transition rate from off state to on state: δ_i (off/on rates)

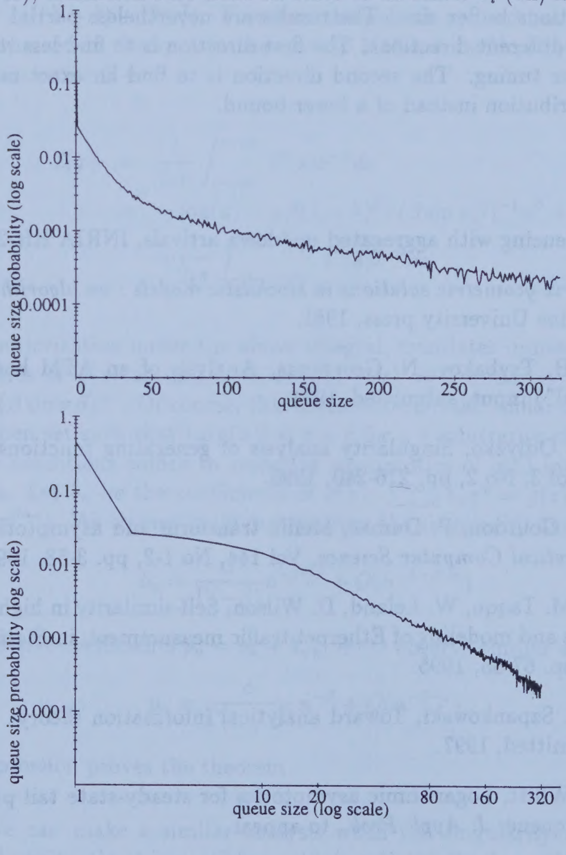


Figure 3: Queue size distribution obtained by simulation (bottom: logarithmic scale), on/off aggregation

[10] W. Szpankowski, On Certain Recurrences Arising in Universal Coding, preprint.

[11] W. Szpankowski, Techniques of the Average Case Analysis of Algorithms, in *Handbook on Algorithms and Theory of Computation* (Ed. M. Atallah), CRC 1997.

Fractal Network Traffic Modeling: Past, Present, and Future

Bo Ryu

Hughes Research Laboratories
3011 Malibu Canyon Rd., Malibu, CA 90265
Tel: (310) 317-5487, Fax: (310) 317-5695
ryu@isl.hrl.hac.com

Abstract

The discovery of a fractal phenomenon in computer network and compressed packet video traffic has added a new dimension to teletraffic modeling and analysis. But some of the recent research effort on the modeling and analysis of the fractal phenomena have generated results that are seemingly confusing or contradictory. The goal of this study is to correctly digest the recent contributions, examine the current status, and lay out research directions of teletraffic research in the area of fractal network traffic modeling and analysis. By doing so, we hope to provide a unified approach to the understanding of the role and implications of fractal behavior in traffic engineering for emerging networks such as B-ISDN and Integrated Services Internet.

1 Introduction

No one would argue against the claim that the establishment and discovery of a fractal phenomenon in computer network [15, 22, 14] and compressed packet video traffic [9, 2] has added a new dimension to teletraffic modeling and analysis. Indeed, there has been a tremendous amount of interest in fractal network traffic, and as a result, recent years have witnessed a flurry of research devoted to gaining insights into and understanding of the impact of this seemingly ubiquitous behavior on network dimensioning and traffic engineering.

It is worth discussing why there is so much attention to the fractal behavior of network traffic. First, the notion of fractal can often explain complex random phenomena such as Internet traffic dynamics in a surprisingly simple way. Fractal stochastic models with only a handful of parameters capture various fractal behavior such as long-range dependence (LRD), slowly decaying variance, and $1/f$ noise in an intuitive way [26]. This is in line with the strength of fractal geometry over Euclidian geometry in terms of describing natural phenomenon, objects and shapes [18]. Second, most fractal behavior is difficult, if not impossible, to capture with traditional Markov models [1, 24]. It is not uncommon that a Markovian model requires tens or even hundreds of parameters and not-intuitive fitting procedure to mimic fractal behavior over a wide range of time scales.

Third, and perhaps most importantly, some of the early studies indicate that the tail probability of a fractal source decays much more slowly than the exponential decay that Markovian models predict. Considering that most performance analysis techniques in teletraffic modeling and analysis have been built upon Markovian models, this deviation has strongly challenged existing teletraffic models and techniques.

Despite the significant attention the second-order self-similarity has enjoyed recently, a clear picture of its nature and importance has not yet emerged from a network performance point of view [2, 8, 11, 27]. The goal of this study is to correctly digest the recent contributions, examine the current status, and lay out research directions of teletraffic research in the area of fractal network traffic modeling and analysis. By doing so, we hope to provide a unified approach to the understanding of the role and implications of fractal behavior in traffic engineering for emerging networks such as B-ISDN and Integrated Services Internet.

2 Past

As discussed in [25], we find it much easier to understand the recent results on the impact of fractal characteristics of network traffic on traffic engineering if we divide fractal behavior into two types: *isolated* fractal behavior and *interaction-induced* fractal behavior. A traffic source exhibits *isolated* fractal behavior over a wide range of time and frequency scales *without any interaction with network or other sources.* A good example of a source-level fractal source is the Star Wars VBR video sequence which was found to exhibit LRD [9]. *Interaction-induced* fractal behavior, in contrast, is observed *as a result of interaction between network actions (e.g., packet dropping), protocol actions (e.g., changing transmission rate or retransmitting lost information), and user actions (e.g., generating connections with diverse workload at various times).* For example, the fractal nature of packet networks (Ethernet and Internet) reported in various measurement studies [15, 14, 22] appears to be caused by these interactions.

2.1 Isolated fractal behavior

The importance of source-level fractal behavior lies in the fact that it is directly relevant to network dimensioning through resource allocation and quality of service (QoS) guarantees. Until the LRD of various VBR video traces [2, 9, 12] has been established, the majority of VBR video traffic models are based on short-range dependent processes such as Markov-modulated Poisson processes [10, 29]. However, new theoretical queueing results with LRD input traffic reported in [6, 17, 20, 21] were more than sufficient to caution researchers and practitioners who depended on Markov-based performance models. In addition, experimental queueing analysis reported in [8, 26] and the spectral analysis of queueing systems performed in [16] show convincing results that LRD has a dominant impact on resource allocation and QoS guarantees.

Different, and perhaps opposing viewpoints about the impact of the LRD property also appear in the literature, however. Heyman and Lakshman [11] and Elwalid *et al* [7] analyzed several traces of VBR videoconferencing traffic and modeled them using the first-order Discrete Autoregressive processes [DAR(1)] by matching the marginal distribution of the frame size and only the lag-1 correlation. Despite the fact that these traces were shown to possess long-range dependence by Beran *et al* [2], Elwalid *et al* showed that for practical ranges of buffer size and cell loss probability (CLP)[1], the DAR(1) model, which is a Markov (short-range dependent, SRD) process, results in accurate prediction of CLPs for the purpose of real-time connection acceptance control.

Ryu and Elwalid [27] explain these seemingly contradicting results using the notion of *Critical Time Scale (CTS)*. Broadly speaking, CTS implies that traffic behavior of a source beyond this time scale does not contribute to QoS such as cell loss probability. It also implies that whether a source exhibits LRD or not, a short-range dependent model would be sufficient to predict QoS of the source when the CTS is small. For real-time VBR video sources it is very likely that the buffer size will be kept relatively small to meet stringent delay jitter requirement, and the required cell loss probability will be very small (say, less than 10^{-6}). The results in [27] show that the CTS tend to be small under these QoS requirements. This explains why the DAR(1) process has been very effective for predicting the cell loss probability of LRD traces in [7]. Under the maximum delay requirement at 10 – 20 msec and cell loss probability requirement of 10^{-6}, correlations between frames that are more than 1 frame apart had little contribution to buffer overflows.

Then, how does one explain the opposite results, such as in [8]? Our intelligent guess is that the system parameters chosen in those studies yield a very large CTS in which long-term correlations are significantly more dominant than short-term correlations. For example, let us closely look at [8, Fig. 6] which shows that capturing long-term correlations yield the best prediction of the buffer behavior of the original trace. Note that the loss probability at the buffer size of 0 is larger than 10%. This implies that the system operates at a fairly high utilization. Since decreasing link capacity (thus increasing utilization) increases CTS of the system [27], it is reasonable to argue that the high utilization chosen for that simulation yields very high value of CTS. Consequently, short-term correlations have little impact on buffer behavior at such an operating point.

2.2 Interaction-induced fractal behavior

The manifestation of the fractal nature of data network traffic has not only been established in Ethernet traffic [15] but also in wide-area-network traffic [14, 22]. With mathematically sound analysis of these measurement studies that the network-level fractal behavior appears to be a ubiquitous phenomenon in data networks such as the Internet, tremendous effort has been devoted to revealing the causes of this network-level fractal behavior. Indeed, explaining the network-level fractal behavior has been a principal research topic during the past two to three years.

Willinger *et al* [31] employs the LRD process constructed by Taqqu and Levy [30] to explain the fractal behavior of Ethernet traffic. This particular LRD process is obtained by superposing i.i.d. ON/OFF (or renewal-reward) processes where both ON and OFF periods are i.i.d. with heavy-tailed distribution. By analyzing the traffic generated by each Ethernet host, Willinger *et al* found that each host generates traffic closely following the heavy-tailed ON/OFF behavior. By connecting this finding to the LRD process of Taqqu and Levy, the fractal behavior of Ethernet traffic is established.

While this approach is based on the mathematically well-constructed LRD model, it lacks some practical factors. First, the cause of heavy-tailed behavior of each host has not been fully investigated. It could be caused by congestion inside the Internet, together with adaptation to network congestion by transport protocols such as TCP, by the application-level traffic which tends to have heavy-tailed file size distributions, or by both. Second, no connection has been made between user behavior (e.g., connection arrival process) and the fractal nature of Ethernet traffic. As will be shown later, user behavior plays an important role determining the factors contributing to the network-level fractal phenomenon.

Crovella and Bestavros [4] suggest Cox's immigration/death model (also known as $M/G/\infty$ model) as a means of explaining the self-similarity of World Wide Web (www) traffic. They found that the distribution of file sizes at a web server obeys a heavy-tailed form. Assuming that the request arrival process to this web server follows a Poisson process, one can readily identify the similarity between the $M/G/\infty$ model and the self-similarity of web traffic.

This approach also suffers from several practical issues. First, as will be shown later, web request process do not always obey a Poisson form. Second, the $M/G/\infty$ model does not model the fluctuation of transmission rate during each connection; the model assumes constant transmission rate during the entire connection. Third, the $M/G/\infty$ model also implies that the connection duration and file size to be transfered is directly proportional. This has not been strongly established, and requires a further study on its validity.

A more practical and heuristic approach to explaining the network-level fractal behavior is reported by Klivansky *et al* [14]. They conducted a very interesting experiment to determine which factors contribute to the network-level fractal behavior. The factors considered include connection arrival process, connection volume (bytes or packets per connection), throughput of each connection, and packet arrivals within each connection. They employ so called the "nested factors" to quantitatively determine which factors contribute to the Hurst parameter, and conclude that the heavy-tailed joint distribution of TCP connection volumes (sizes) and connection transmission rates are the key factors.

Ryu and Lowen [25] employ the Fractal-Shot-Noise-Driven Poisson Point Process (FSNDP) to explain the fractal behavior of TCP traffic. The FSNDP model, one of many fractal-modulated Poisson processes constructed and analyzed in [26], consists of the first-stage homogeneous Poisson process $dM(t)$ and the second-stage Poisson process connected via a linear filter $h(t)$. The process $dM(t)$ captures the TCP connection arrival process which, by definition, must follow a Poisson form. In this sense, the FSNDP suffers the same limitation as the $M/G/\infty$ model; only the Poisson-obeying TCP connection arrival process can be captured. However, this model can additionally capture the connection volume (area of $h(t)$) and transmission rate fluctuation (height and decaying form of $h(t)$), making it more versatile and less heuristic than [14]. Later, we discuss a generalized version of the FSNDP which can capture (almost) an arbitrary connection arrival process.

3 Present

3.1 Isolated fractal behavior

One area that researchers are continuously working on is developing new queueing theories with fractal input processes (or heavy-tailed service requirements). Since the fractal behavior of a source is defined in an asymptotic sense, the queueing results for fractal input traffic also tend to be in an asymptotic form [5, 20, 21, 17]. Recently, new results are being reported on computing the queue size distribution with either self-similar processes or heavy-tailed service requirements over the entire queue size (not just the tail) [13, 23]. From a queueing theory point of view, it is expected that these new results will be quite useful for various queueing-oriented problems.

3.2 Interaction-induced fractal behavior

We briefly discuss the generalized FSNDP (G-FSNDP) and its application for explaining the causes for TCP traffic self-similarity. As mentioned earlier, the G-FSNDP model has an arbitrary input point process $dM(t)$ instead of homogeneous Poisson process as in the FSNDP, and thus can capture almost any connection arrival process under mild conditions. This provides a unique strength of this model over other approaches, since the connection arrival processes are not necessarily Poisson. In fact, the connection arrival process itself can be fractal. Consequently, this G-FSNDP model reveal three cases which yield network-level fractal behavior: (i) connection arrival process is fractal, (ii) filter function $h(t)$ is fractal (due to network action, protocol action, and connection volume), or (iii) both. Currently, work is under way to apply the G-FSNDP model to characterize a fractal packet trace.

An example of fractal connection arrival process is given below. We have analyzed a four-hour long web request data recently collected at UC Berkeley. [1] The data is not only up-to-date (less than a year old), but also sufficiently large (100k requests) that statistical analysis with sufficient accuracy becomes possible.

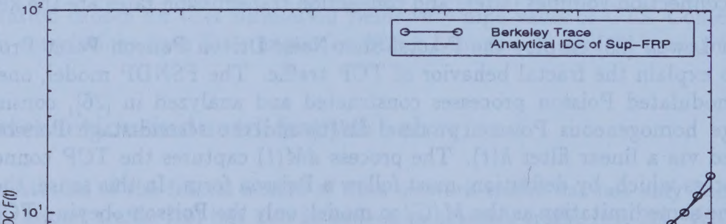


Figure 1: Index of Dispersion for Counts (IDC) of Berkeley trace and the matched Sup-FRP model.

Figure 1 shows an estimated *Index of Dispersion for Counts (IDC)* curve of the Berkeley web request trace and an analytical IDC curve of the matched Sup-FRP model [28]. The fitting procedure is described in [26]. The straight line over the log-log plot over the time scale in the range of 1 – 100 sec is evident, indicating fractal behavior over this range of time scales. Furthermore, the matched Sup-FRP model closely approximates

---

[1] This trace is available at http://www.acm.org/sigcomm/ITA.

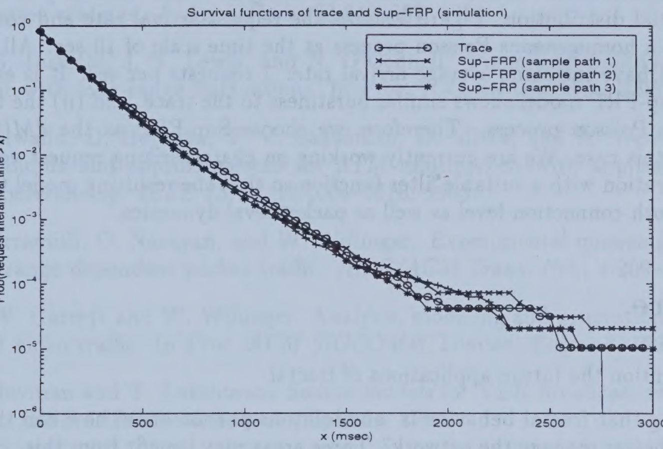


Figure 2: Estimated survival functions of the inter-request times of the Berkeley trace and the matched Sup-FRP model (three sample paths shown).

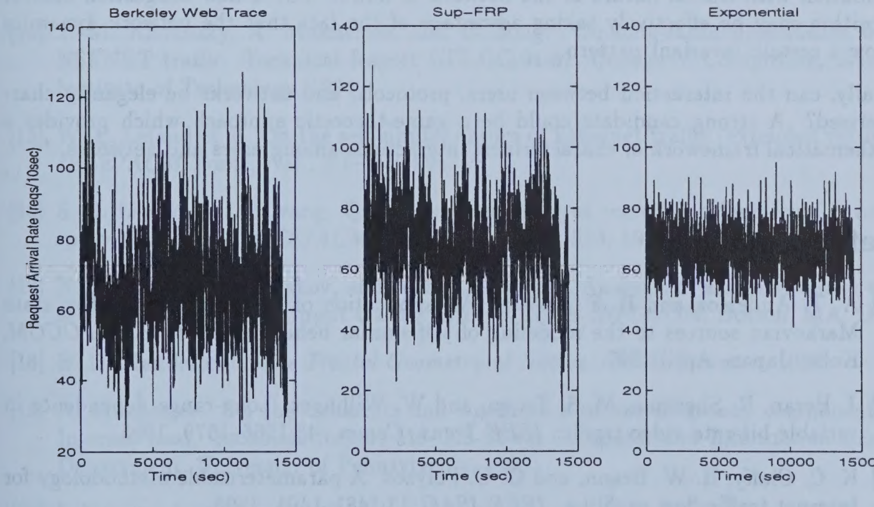


Figure 3: Comparison of request arrival rates of the trace, Sup-FRP model, and homogeneous Poisson model at the time scale of 10 sec. They all have the same average rate of 7 requests per sec.

not only the IDC of the trace, but also the marginal distribution of inter-request times. The estimated survival function of the trace in Fig. 2 indicates exponential distribution of inter-request times of the trace (note the logarithmic scale in the y axis). However, they are not uncorrelated as indicated in the IDC figure [2] and thus they are not from i.i.d. exponential distribution. Figure 3 shows the request arrival rate and compares to Sup-FRP and a homogeneous Poisson process at the time scale of 10 sec. All the three plots in Fig. 3 have the same average arrival rate: 7 requests per sec. It is easy to see that (i) the Sup-FRP model shows similar burstiness to the trace, and (ii) the trace does not follow the Poisson process. Therefore, we choose Sup-FRP as the $dM(t)$ for the G-FSNDP in this case. We are currently working on characterizing request volume and connection duration with a suitable filter function so that the resulting model G-FSNDP can capture both connection-level as well as packet-level dynamics.

4 Future

We briefly mention the future applications of fractal.

First, assuming that fractal behavior *is* a ubiquitous phenomenon, how can this nature be applied to better manage the network? Three areas may benefit from this: congestion prediction [19], traffic volume prediction, and flow modeling [3]. It seems more and more evident that congestion level of network and traffic fractal characteristics are closely related. by utilizing this relation, one may be able to manage the network more intelligently and efficiently.

Second, can the next generation TCP (TCP-NG) benefit from the fact that network dynamic is fractal? There have been suggestions within the Internet research community about making changes in TCP. [3] A possible candidate is to develop a novel round-trip time estimation with fractal nature of the network in mind. TCP's new congestion control algorithm may be effectively taking advantage of the fact that the network dynamics follow a certain invariant pattern.

Finally, can the interaction between users, protocols, and networks be elegantly characterized? A strong candidate could be a game-theoretic approach which provides a mathematical framework of characterizing interactions among users and networks. [4]

References

[1] A. T. Anderson and B. F. Nielsen. An application of superpositions of two state Markovian sources to the modeling of self-similar behavior. In *Proc. INFOCOM*, Kobe, Japan, April 1997.

[2] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Trans. Comm.*, 43:1566–1579, 1995.

[3] K. C. Claffy, H.-W. Braun, and G. C. Polyzos. A parameterizable methodology for Internet traffic flow profiling. *IEEE JSAC*, 13:1481–1494, 1995.

---

[2] A time series of i.i.d. exponentially distributed interarrival times yields the IDC value of 1 regardless of the time window.

[3] End-to-end interest e-mail discussions.

[4] See http://comet.ctr.columbia.edu for related work.

[4] Mark Crovella and Azer Bestavros. Explaining World Wide Web traffic self-similarity. Technical Report TR-95-015, Boston University, CS Dept, August 1995.

[5] N. G. Duffield. Economies of scale in queues with sources having power-law large deviation scalings. *J. Appl. Prob.*, 33:840–857, 1996.

[6] N. G. Duffield, J. T. Lewis, and N. O'Connell. Predicting Quality of Service for traffic with long-range fluctuations. In *Proc. ICC*, Seattle, WA, 1995.

[7] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE JSAC*, 13:1004–1016, 1995.

[8] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Net.*, 4:209–223, 1996.

[9] M. W. Garrett and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proc. ACM SIGCOMM*, London, England, 1994.

[10] D. Heyman and T. Lakshman. Source models for VBR broadcast-video traffic. In *Proc. IEEE INFOCOM*, Toronto, Canada, 1994.

[11] D. P. Heyman and T. V. Lakshman. What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Trans. Net.*, 4:301–317, 1996.

[12] P. Jelenkovic, A. Lazar, and N. Semret. Multiple time scales and subexponentiality in MPEG video streams. In *Proc. Broadband Comm.*, Montréal, Canada, 1996.

[13] P. R. Jelenkovic and A. A. Lazar. Multiplexing on-off sources with subexponential on periods: Part i. In *IEEE INFOCOM*, Kobe, Japan, 1997.

[14] S. M. Klivansky, A. Mukherjee, and C. Song. On long-rante dependence of in NSFNET traffic. Technical Report GIT-CC-94-61, College of Computing, Georgia Institute of Technology, 1994.

[15] W. E. Leland et al. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. Net.*, 2:1–15, 1994.

[16] S. Q. Li and C. L. Hwang. Queue response to input correlation functions: Discrete spectral analysis. *IEEE/ACM Trans. Net.*, 1:522–533, 1993.

[17] N. Likhanov, B. Tsybakov, and N. D. Georganas. Analysis of an ATM buffer with self-similar ("fractal") input traffic. In *Proc. IEEE INFOCOM*, Boston, MA, 1995.

[18] B. B. Mandelbrot. *The Fractal Geometry of Nature.* W. H. Freeman, 1982.

[19] A. Mukherjee. On the dynamics and significance of low frequency components of Internet load. Technical Report MS-CIS-92-83, Computer and Information Science Department, University of Pennsylvania, 1992.

[20] I. Norros. A storage model with self-similar input. *Queueing Systems*, 16:387–396, 1994.

[21] M. Parulekar and A. M. Makowski. Buffer overflow probabilities for a multiplexer with self-similar input. In *Proc. IEEE INFOCOM*, San Francisco, CA, 1996.

[22] V. Paxson and S. Floyd. Wide Area Traffic: the failure of Poisson modeling. *IEEE Trans. Net.*, 3:226–244, 1995.

[23] M Raughan, D. Veitch, and M. Rumsewicz. Computing queue-length distributions for power-law queues, 1997. preprint.

[24] S. Robert and J.-Y. LeBoudec. Can Self-Similar traffic be modeled by Markovian processes? In B. Plattner, editor, *Lecture Notes in Comp. Sci. (Proc. Int'l Zurich Seminar on Dig. Comm.)*, volume 1044. Springer-Verlag, 1996.

[25] B. Ryu and S. Lowen. Point process approaches for modeling and analysis of self-similar traffic: Part ii – applications. In *Proc. 5th Int'l Conf. on Telecommunications Systems, Modeling, and Analysis*, Nashville, TN, 1997.

[26] B. Ryu and S. Lowen. Point process models for self-similar network traffic, with applications. To appear in *Stochastic Models*, 1997.

[27] B. K. Ryu and A. Elwalid. The importance of Long-Range Dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *Proc. ACM SIGCOMM*, San Francisco, CA, 1996.

[28] B. K. Ryu and S. B. Lowen. Point process approaches to the modeling and analysis of self-similar traffic: Part I - Model construction. In *Proc. IEEE INFOCOM*, San Francisco, CA, 1996.

[29] P. Skelly. *Characterization and control of variable bit rate video in broadband networks.* PhD thesis, Columbia University, 1994.

[30] M. S. Taqqu and J. B. Levy. Using renewal processes to generate long-range dependence and high variability. In E. Eberlein and M. S. Taqqu, editors, *Dependence in Probability and Statistics*, volume 11, pages 73–89. Birkhauser, Boston, MA, 1986.

[31] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through high-variability: Statistical analysis of Eternet LAN traffic at the source level. In *Proc. ACM SIGCOMM*, Cambridge, MA, 1995.

An empirical Internet traffic study

J. Aracil, R. Edell, P. Varaiya
Dept. Electrical Engineering and Computer Sciences
University of California, Berkeley

Abstract

One day's worth of WAN Telnet and FTP connections from the Berkeley campus is analyzed. The main conclusions of the analysis are these. A wide diversity of use offers opportunities to reduce bandwidth by traffic shaping and shifting. Second, connection size and duration are heavy-tailed. Third, Telnet packet arrivals are heavy-tailed, but FTP packet arrivals are not. Both Telnet and FTP connection births have exponential interarrivals, although they are not correlated. Lastly, the TCP flow control mechanism significantly affects TCP heavy-tailed behavior.

1 Introduction

This study is part of a growing body of empirical work on wide area Internet traffic. The main conclusions accord are in accord with those reported in [6, 5]. Those conclusions can be summarized as follows.

There is a wide diversity of use that offers opportunities to reduce bandwidth from traffic shaping and shifting. Second, connection size and duration are heavy-tailed. Third, Telnet packet arrivals are heavy-tailed, but FTP packet arrivals are not. Both Telnet and FTP connection births have exponential interarrivals, although they are not correlated. Lastly, the TCP flow control mechanism significantly affects TCP heavy-tailed behavior. In addition, the data suggest that multiplexing Telnet connections may only provide small gain.

2 Empirical setting and diversity of use

We analyze four days worth of IP traffic between UCB (U.C. Berkeley) and the rest of the world over the period January 17-20, 1995. The recorded data comprises IP packet header information with a 50 ns time stamp. In total 439M packets (89 Gbytes) were exchanged, at 0.1 s peak rates ranging from 2K to 6K packets per second or 0.8M to 1.1M bytes per second. This traffic load originates from 6,300 hosts (out of a total of 22,000 hosts) on campus.

TCP connections account for 92% of all data transferred. At any given time there were between 600 and 1,200 active connections, and between 41 and 79 new connections per second. (So the mean connection duration is about 20 seconds.)

By protocol, the TCP data transferred divides as follows: ftp (33%), nntp (24%), www (16%), smtp (9%), shell (4%), telnet (3%), X (3%), other (8%). We analyze FTP and Telnet traffic in greater detail.

There is a wide diversity among users. Analysis of individual users in Electrical Engineering & Computer Sciences reveals that 10% of users account for 80% of WAN traffic volume and 20% of users account for more than 90% of volume. (This is similar to the so-called 80-20 rule for telephony.) This diversity has usage-based pricing implications. (Like most institutions, the Berkeley campus does not at present have such charges.)

The aggregate campus-wide traffic is bursty, and that burstiness could be exploited to reduce the peak bandwidth in the link connecting the campus to the Internet backbone, by traffic-shaping and traffic-shifting. The peak 1 second bandwidth is 76% of the peak 0.1 second bandwidth. This means that if traffic can be delayed up to 1 second (requiring a 250K byte buffer), the peak bandwidth can be reduced by 24%. On the other hand suppose we somewhat arbitrarily divide all traffic into delay-sensitive traffic (ftp, www, shell, telnet, X) and delay-tolerant traffic (smtp,nntp). Next we shift the latter traffic to those times where the traffic is of below-average intensity. This shift reduces the peak 20-minute average intensity by 23%. Such shift would of course require much larger buffers or incentives that cause users to shift the traffic. For further detail, see [3]

3 Connection duration, size

We study Telnet, FTP control, and FTP data connections. The duration of these TCP connections is defined to be the time between SYN (connection start) and FIN (connection end) packets.[1] Size is the number of bytes transmitted in the connection.

We analyze one day's data amounting to 13K Telenet, 23K FTP control and 52K FTP data connections. It turns out that the marginal distributions of size and durations are all heavy-tailed, i.e, the tails fit a power law:

$$P(X > x) \approx x^{-\alpha},$$

for large x. We fit a straight line to the plot of the log-log empirical distributions of size and duration. The slope of the fitted straight line is the estimate of α given in the table below:

Connection type	Estimate of α (size)	Estimate of α (duration)
Telnet	2.14	1.18
FTP Control	2.5	2.25
FTP Data	1.92	1.28

Heavy tail estimates, 1-day's data

The heavy-tailed behavior of FTP and Telnet connection durations and sizes has been previously well-established [6].

---

[1] The empirical situation is more complex. TCP connection initiation and termination don't always follow the textbook description. There are connections which end without FIN or start without any activity. Empirically, end is defined in terms of a FIN or period of inactivity.

4 Telnet and FTP data packet birth process

Within a connection packets arrive randomly. We take 25 Telnet and 25 FTP data connections and plot the log-log survival functions and the Hill estimate [4] of the packet inter-arrival times. The plots are in Figure 1.

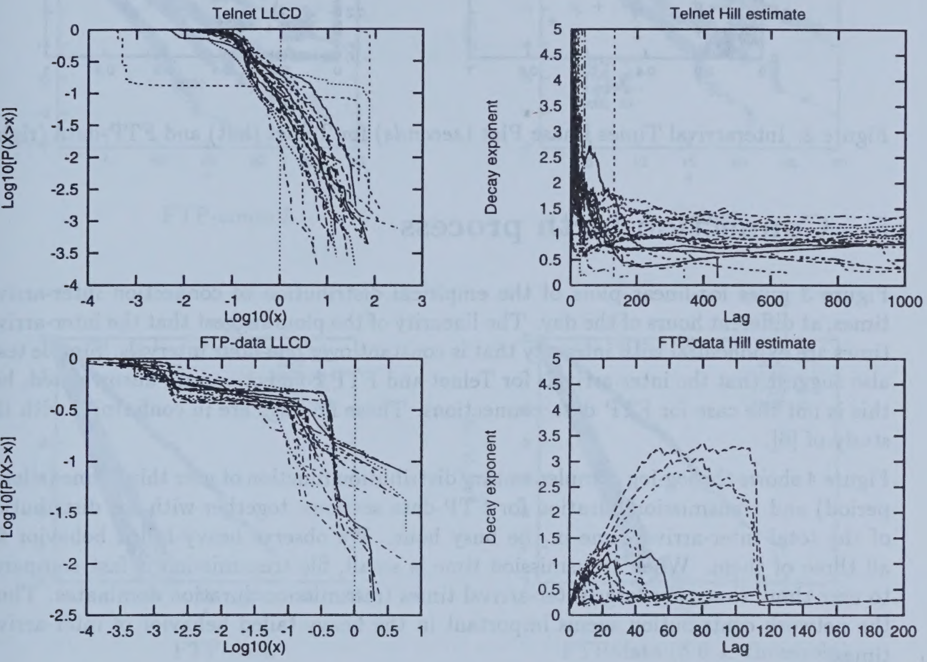


Figure 1: Log-log survival function and Hill estimate for Telnet (above) and FTP data (below) packet inter-arrival times; x = bytes.

The Hill estimator either stabilizes around the tail decay exponent, or it tends to decay if the distribution is not heavy-tailed. The figure suggests heavy-tailed behavior for Telnet packet inter-arrivals ($\alpha \approx 1$) but not for FTP data packets within the same connection.

We perform a simple serial correlation test ([2, chapter 6]) of our Telnet and FTP data connection sample. At a level of significance of 5% we find that the 42.7% of the Telnet connections and the 35.4% of the FTP-data connections do not have independent packet arrivals. To test for appropriateness of ON-OFF models for FTP data connections we perform the same serial correlation test to filtered versions of the packet arrival process. We define OFF period as the period of time for which no consecutive packet arrivals take place separated by less than a certain threshold. We consider thresholds of 0.5, 1.0, 1.5 and 2.0 seconds and do not perceive significant changes. See [6] for similar tests.

Some understanding of the lack of correlation is gained from the deterministic features in the "phase" plot of successive inter-arrival times $X(i)$ within the same connection, Figure 2. The clusters around the points (0.2, 0.4) and (0.4, 0.4) (seconds) may be due to TCP time granularity and the effect of TCP window mechanism.

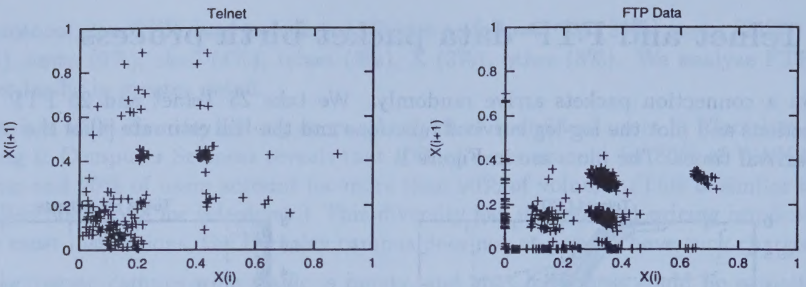


Figure 2: Interarrival Times Phase Plot (*seconds*) for Telnet (left) and FTP-data (right)

5 Connection birth process

Figure 3 gives log-linear plots of the empirical distribution of connection inter-arrival times, at different hours of the day. The linearity of the plots suggest that the inter-arrival times are exponential with intensity that is constant over one-hour intervals. Simple tests also suggest that the inter-arrivals for Telnet and FTP connections are uncorrelated, but this is not the case for FTP data connections. These findings are in conformity with the study of [6].

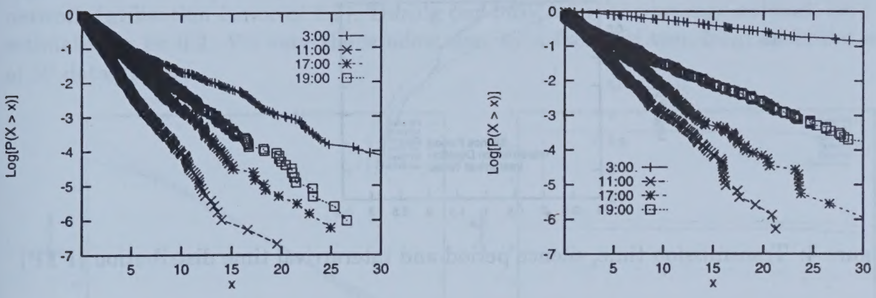
Figure 4 shows the log-log complementary distribution function of user think time (silence period) and transmission duration for FTP-data sessions, together with the distribution of the total inter-arrival time in the busy hour. We observe heavy-tailed behavior for all three of them. When transmission time is small, file transmission is fast compared to user think time. For large inter-arrival times transmission duration dominates. Thus, the network contribution seems important in the heavy-tailed behavior of inter-arrival times.

6 Multiplexing gain

The marginal distribution of bytes per interval of time of several connections can suggest the extent of statistical multiplexing gain. Let X_n^M denote the number of bytes in the nth 10 ms. interval coming from a multiplex of M connections, namely,

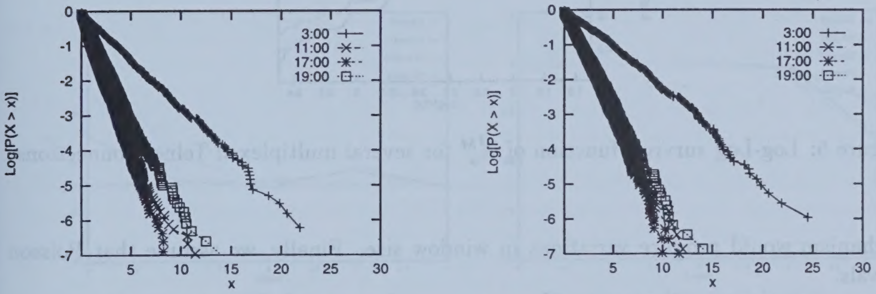
$$X_n^M = \sum_{i=1}^{M} X_n^i$$

where X_n^i represents the number of bytes in the nth interval that connection i generates. Figures 5 and 6 plots the empirical log-log survival functions for several Telnet and FTP data connections respectively. Figure 5 suggests that for the Telnet case these distributions are heavy-tailed, and the decay exponent does not change with the number of multiplexed connections. The exponent α is in the range $1 < \alpha < 2$, indicating infinite variance. However, for FTP data connections, Figure 6, the survival function is staircase shaped, suggesting bursty, non-continuous transmission. Also, increasing the number of multiplexed connections has a smoothing effect.



FTP-control

Telnet



FTP-data

FTP-data (5.0 s. threshold)

Figure 3: Log-linear survival function for connection inter-arrival times.

7 Network influence

Transmission time and file sizes are not related linearly since TCP feedback dynamics make transmission time grow larger. The effect of window size can be quite striking. We propose a simulation model for TCP/IP connections that consists of a delayed feedback queue to simulate a link between source and destination. The source starts new connections which consist of datagram batch arrivals with Pareto distribution in bytes to reflect the heavy tail of the FTP data size distribution. We suppose that a window size worth of datagrams is transmitted together and therefore the service rate in the server is constant and equals the negotiated window size divided by the link speed. Each window-size worth of datagrams triggers another window-size worth of datagrams (after the delay) until the whole batch gets transmitted.

This simplified behavior is common to window-based congestion flow control algorithms. The roundtrip delay between source and destination remains constant within connection duration. Window size is also assumed to be constant, although the TCP Slow Start

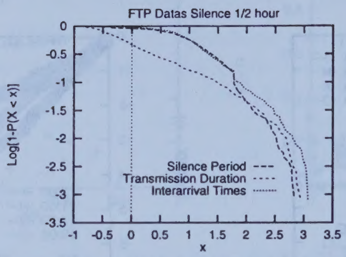


Figure 4: Transmission time, silence period and interarrival time distribution (FTP)

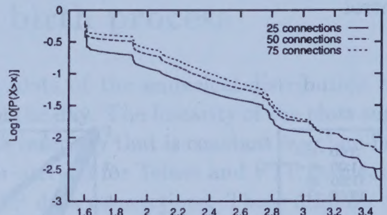


Figure 5: Log-Log survival function of X_n^M for several multiplexed Telnet connections

mechanism would produce variations in window size. Finally, we assume that Poisson arrivals.

The link speed is assumed to be 10 Mbps (Ethernet rates) and the α parameter is fitted to the empirical FTP data size distribution parameter value of 1.28. We consider roundtrip times of 0.1, 0.25 and 1.0 seconds and we vary the value of the window size from 50 to 100 datagrams at the same utilization factor.

We measure packet arrivals in 1 second intervals at the output of the server. Assuming such server represents a transmission link we are measuring network traffic as in our real measurement scenario. We produce variance-aggregation level log-log plots of the packet counting process and we also calculate the Whittle estimator [1] for the parameter H

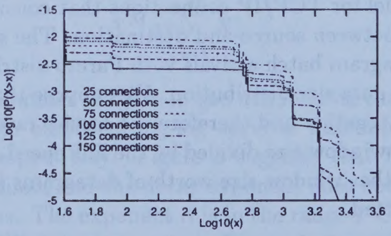
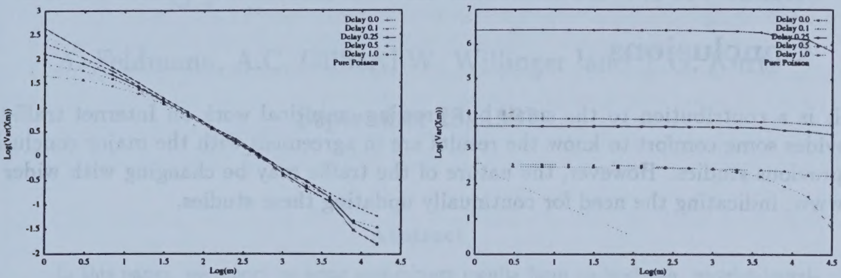


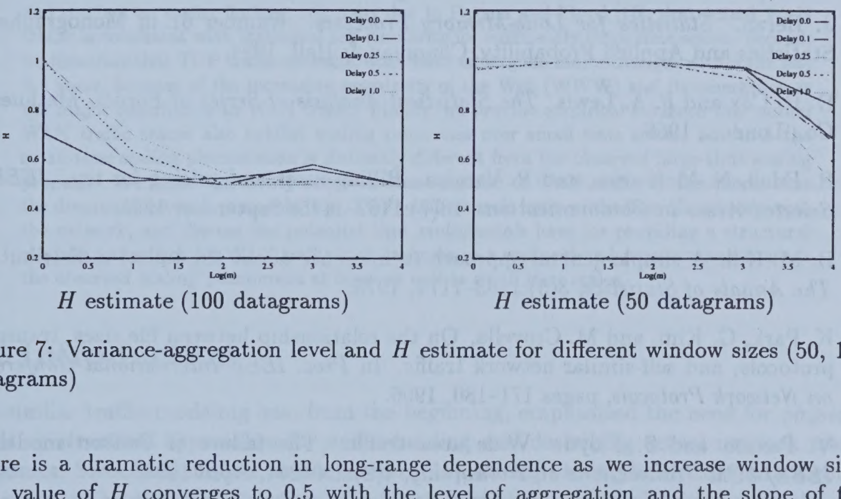
Figure 6: Log-Log survival function of X_n^M for several multiplexed FTP-data connections

using the aggregation method with 95% confidence intervals.

Figure 7 shows a summary of results of the simulation experiment with a fairly loaded network (utilization factor of 0.7). During our busy hour, the average network load was estimated to be 0.3. We vary the window size, by a factor of two, from an initial value of 50 datagrams.



Variance vs. agg. level (100 datagrams) Variance vs. agg level (50 datagrams)



H estimate (100 datagrams) H estimate (50 datagrams)

Figure 7: Variance-aggregation level and H estimate for different window sizes (50, 100 datagrams)

There is a dramatic reduction in long-range dependence as we increase window size. The value of H converges to 0.5 with the level of aggregation and the slope of the variance-aggregation level line tends to -1, approaching independent behavior in the packet counting process with a window size of 100 datagrams. However, the H estimate is close to 0.9 with a window size of 50 datagrams. These results match our intuition: the delay effect clearly dominates over bandwidth limitations, so that the smaller the window size the more *window roundtrips* have to be performed between origin and destination hosts in order to transmit the wohle datagram batch. Each roundtrip introduces a delay that spreads the batch over several time slots, creating long-range dependence. The service time is always an order of magnitude below the delay considering a 10 Mbps. link, so that the contribution of network load is not as important as the delay contribution. See also [5].

This relates to the model presented in [7], where a second-order self-similar process is constructed with the counting process of Poisson arriving bursts with heavy-tailed distribution. Small values of the window size will make the file transmission time last longer, producing heavy-tail behavior and thus resembling the model presented in [7]. For large window sizes, the traffic at the output of the server resembles the Poisson process at the input of the queue, whereas in the small window case the duration of the multiplexed traffic bursts becomes heavy-tailed and, therefore, we are faced with a self-similar process.

8 Conclusions

This is a contribution to the small but growing empirical work on Internet traffic. It provides some comfort to know the results are in agreement with the major conclusions of previous studies. However, the nature of the traffic may be changing with wider use fo www, indicating the need for continually updating these studies.

References

[1] J. Beran. *Statistics for Long-Memory Processes.* Number 61 in Monographs on Statistics and Applied Probability. Chapman & Hall, 1994.

[2] D. R. Cox and P. A. Lewis. *The Statistical Analysis of Series of Events.* Methuen & Co., London, 1966.

[3] R. Edell, N. McKeown, and P. Varaiya. Billing users and pricing for tcp. *IEEE J. Selected Areas in Communications*, 13(7):1162–1175, September 1995.

[4] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[5] K. Park, G. Kim, and M. Crovella. On the relationship between file sizes, transport protocols, and self-similar network traffic. In *Proc. IEEE International Conference on Network Protocols*, pages 171–180, 1996.

[6] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 4(2):226–244, April 1996.

[7] B. Tsybakov and N. D. Georganas. On self-similar traffic in ATM queues: Definitions, overflow probability bound and cell delay distribution. *IEEE/ACM Transactions on Networking*, 5(3):397–409, June 1997.

Looking behind and beyond self-similarity: On scaling phenomena in measured WAN traffic[*]

A. Feldmann, A.C. Gilbert[†], W. Willinger [‡]and T.G. Kurtz

September 28, 1997

Abstract

In this paper, we report on some preliminary results from an in-depth, wavelet-based analysis of a set of high-quality, packet-level traffic measurements, collected over the last 6-7 years from a number of different working wide-area networks (WANs). We first validate and confirm an earlier finding, originally due to Paxson and Floyd [10], that actual WAN traffic is consistent with statistical self-similarity for sufficiently large time scales. Second, we illustrate that TCP traffic characteristics have undergone major changes within the last 3-4 years, because of the increasing popularity of the Web (WWW) and its emergence as the major contributor to WAN traffic. Finally, we provide empirical evidence that actual WAN traffic traces also exhibit scaling properties over small time scales, but that the small-time scaling phenomenon is distinctly different from the observed large-time scaling property. We relate this newly observed characteristic of WAN traffic to the effects that the dominant network protocols (e.g., TCP) and controls have on the flow of packets across the network, and discuss the potential that *multifractals* have for providing a structural modeling approach for WAN traffic that captures in a compact and parsimonious manner the observed scaling phenomena at large as well as small time scales.

1 Introduction

Self-similar traffic modeling has, from the beginning, emphasized the need for *physical-based* or *structural* approaches for understanding and describing actual network traffic dynamics. Structural models, as discussed in [13], attempt to implicitly take into account the complex hierarchical structure of modern computer communications networks and the intertwined networking mechanisms that determine the nature of the traffic which these networks carry. These structural models are different from the "black box" or "operational" traffic models which were popular in the past.

The main objective of this paper is to improve our current understanding of the dynamic nature of traffic carried over wide-area networks (WANs) such as the Internet and to outline new structural modeling approaches for WAN traffic that are consistent with actual measurements. In particular, we (i) identify distinct scaling regions in aggregate packet-level WAN traces, one for small time scales (a few hundreds of milliseconds and below) and one for large time scales (beyond a few hundreds of milliseconds); (ii) demonstrate the changing nature of WAN traffic characteristics at the transport (i.e., TCP/IP) level over time; (iii) provide empirical evidence for the presence of application-layer WAN traffic characteristics, which allows us to adopt a structural modeling approach that explains the empirically observed large-time scaling phenomenon in aggregate WAN traffic; and (iv) discuss the feasibility of a structural modeling approach for WAN traffic that is capable of capturing the large-time as well as small-time scaling properties in a compact and parsimonious manner. To accomplish these goals, we rely on a unique set of high-quality, packet-level WAN traffic measurements, collected over the past six to seven years from a number of different working WANs. Our in-depth analysis of these WAN traffic traces is based on a wavelet-based technique recently suggested and developed by Abry and Veitch [1]. This technique provides a natural and effective tool for investigating scaling properties that may be present in a large data set.

---

[*] A. Feldmann, A. Gilbert, and W. Willinger are with AT&T Labs-Research, Florham Park, NJ 07932-0971, email: {anja, agilbert, walter}@research.att.com. T. Kurtz is with the Center for the Mathematical Sciences, University of Wisconsin at Madison, Madison, WI 53717, email: kurtz@math.wisc.edu.

[†] This Research was supported by the NSF Grant DMS-9705665 at Yale University and AT&T Labs-Research, Florham Park, NJ.

[‡] This research was partially supported by the NSF Grant NCR-9628067 at the University of California at Santa Cruz.

There have been significant advances in the recent past in structural modeling approaches for local area network (LAN) traffic as opposed to WAN traffic. In fact, the empirically observed self-similar or fractal nature of LAN traffic at the aggregate or macroscopic level (i.e., generated by all active hosts on the network; see for example [9]) has recently resulted in structural models that are based on self-similar processes, have a physical meaning in the LAN context, and provide fundamental and new insights into how individual network connections behave. However, WANs differ from LANs in a number of fundamental ways, and these differences make structural modeling of WAN traffic more challenging. For example, because of the problem of latency, WANs often rely on sophisticated transport protocols (e.g., Internet uses predominantly TCP) which, in turn, are bound to introduce additional structure to the flow of packets; such a structure is less obvious in LAN traffic because feedbacks in LANs are essentially instantaneous. Despite these differences, empirical evidence for (asymptotic) self-similarity, i.e., large-time scaling phenomena, also exists for WAN traffic and was first reported by Paxson and Floyd [10]; see also [14]. In Section 2 below, we use a collection of more recent WAN traces and a wavelet-based scaling analysis to validate this previously observed large-time scaling property of WAN traffic. In addition, we identify a new scaling phenomenon for small time scales (below a few hundreds of milliseconds) that has gone unnoticed in previous studies and is distinctly different from that for large time scales. In Section 3, we provide new evidence that structural modeling approaches for WAN traffic (e.g., Cox's construction; see [3]) remain partly valid for today's WAN traffic, even though over the last three to four years, WWW has become the main WAN application and typically makes up a major portion of modern WAN traffic. At the same time, we also demonstrate in Section 3 that mainly due to the Web, the traffic characteristics at the transport layer have undergone significant changes between the pre-WWW days and now, and that these changes clearly reveal the limitations of structural modeling a la Cox. To this end, we confirm here with our latest collection of WAN traffic traces a recent finding by Feldmann [5] who observed self-similar features not only at the packet/bytes level but also at the TCP level. In light of this empirical evidence, we revisit the question of structural modeling approaches for modern WAN traffic and discuss two alternatives. One approach which builds on Cox's approach is based on a construction proposed by Kurtz [8] and allows for very flexible "within-session" traffic patterns. A second, more radical approach to understanding and describing the actual dynamics of modern WAN traffic is based on *multifractals* (for details, see for example, [4, 7] and references therein). We conclude in Section 4 with an intuitive explanation for why multifractals might be an appropriate mathematical technique for gaining a better understanding of the dynamic nature of modern high-speed network traffic.

2 Scaling Properties of Measured WAN Traffic

2.1 MRA and a Wavelet-Based Scaling Analysis

Wavelets provide a tool for time-frequency localization; the wavelet transform divides data into different frequency components and analyzes each component with a resolution matched to its scale. We can use the coefficients of a wavelet decomposition to directly study the scale (or frequency) dependent properties of the data. In particular, applying a recently proposed technique by Abry and Veitch [1], we can use wavelets to investigate the scaling structure in the spectra of self-similar processes which provide statistical models for describing certain aspects of modern high-speed network traffic. Multiresolution analysis (MRA) gives us a natural framework for understanding wavelet bases and their transforms.

The key feature of an MRA is that we can write an approximation of a signal X_t at scale j (with resolution 2^j) as the sum of a coarser approximation at scale $j+1$ (with resolution 2^{j+1}) and the difference between these two approximations. We may iterate this procedure, writing the approximation at scale $j+1$ as a sum of a coarser approximation and the difference. This procedure can be implemented using a pyramidal filter-bank algorithm which has computational cost $O(N)$ for data of length N. We write the *wavelet decomposition* of a signal X as

$$X = \sum_{k\in\mathbf{Z}} \langle X, \phi_{0,k}\rangle\, \phi_{0,k} + \sum_{j\leq 0}\sum_{k\in\mathbf{Z}} \langle X, \psi_{j,k}\rangle\, \psi_{j,k} = \sum_{j\in\mathbf{Z}}\sum_{k\in\mathbf{Z}} \langle X, \psi_{j,k}\rangle\, \psi_{j,k}.$$

We call the inner products $\langle X, \psi_{j,k}\rangle$ of X with the rescaled and translated copies of the wavelet ψ the *wavelet coefficients* $d_{j,k}$ of X. The set of all wavelet coefficients is generally referred to as the *discrete wavelet transform (DWT)* of the signal X. The coefficient $|d_{j,k}|^2$ measures the amount of energy in a signal X about the time $t_0 = 2^j k$ and about the frequency $2^{-j}\lambda_0$, where λ_0 is a reference frequency which depends on the wavelet ψ. Abry and Veitch show that the time average of $|d_{j,k}|^2$ at each scale j is a useful spectral estimator. In fact, if E_j denotes the average of $|d_{j,k}|^2$ at each scale, $E_j = N_j^{-1}\sum_k |d_{j,k}|^2$ (N_j is the number of available wavelet coefficients at scale j), then E_j is a measure of the energy that lies within a given bandwidth 2^{-j} around frequency $2^{-j}\lambda_0$.

Consider now a signal X_t generated by a finite variance, wide-sense stationary, *long-range dependent* process with *Hurst parameter* $H \in (1/2, 1)$. In the frequency domain, long-range dependence is characterized by a spectral density $f(\lambda)$ that exhibits a power-law near the origin; i.e., $f(\lambda)$ has the form

$$f(\lambda) \sim c_f |\lambda|^{1-2H} \quad \text{as } \lambda \to 0, \tag{1}$$

where $c_f = 1/\pi\sigma^2 c_\rho \Gamma(2H-1)\sin(\pi(1-H))$ and $\sigma^2 = \mathrm{Var}(X_t)$. Long-range dependence plays an important role in the study of self-similar processes, where a wide-sense stationary process X is called *exactly self-similar* if for all integers $m > 0$,

$$X = m^{1-H} X^{(m)}. \tag{2}$$

Here, the equality is understood in the sense of finite-dimensional distributions, and the aggregated processes $X^{(m)}$ are defined by $X^{(m)}(k) = m^{-1}\sum_{i=(k-1)m+1}^{i=km} X_i$, $k \geq 1$. *Asymptotic self-similarity* is defined similarly but we only require that the above equality holds in the limit as $m \to \infty$. With this definition, X is long-range dependent if and only if X is (asymptotically) self-similar, and methods for studying the dependence structure of X can thus be exploited for investigating the scaling phenomenon expressed in Equation

(2). To illustrate, we return to the wavelet-based technique proposed by Abry and Veitch [1] and assume that the signal X is such that Equation (1) holds for all frequencies. Then the expectation of E_j is given by

$$\mathbf{E}[E_j] = \int f(\lambda) 2^j |\hat{\psi}(2^j \lambda)|^2 \, d\lambda = c_f |2^{-j} \lambda_0|^{1-2H} \int |\lambda|^{1-2H} |\hat{\psi}(\lambda)|^2 \, d\lambda,$$

where $\hat{\psi}(\lambda)$ is the Fourier transform of $\psi(t)$. Observe that the multiplicative bias of E_j is simply a multiplicative constant which is independent of scale j and results in an effective scaling analysis of a given signal X by plotting $\log_2 E_j$ against scale j and identifying scaling regions, breakpoints and non-scaling behavior. It also yields an unbiased estimator for the Hurst parameter H (for details, see [1]). To insure that the integral $\int |\lambda|^{1-2H} |\hat{\psi}(\lambda)|^2 \, d\lambda$ exists and is finite we must choose a wavelet ψ which has $M > H - 1$ vanishing moments. In other words, the wavelet ψ must be orthogonal to all polynomials of degree less than M; i.e., $\int t^l \psi(t) \, dt = 0 \quad \text{for } l = 0, \ldots, M-1$. The number of vanishing moments controls the order of the zero in the Fourier transform of ψ about $\lambda = 0$. If M is large enough, the behavior of $|\hat{\psi}(\lambda)|^2$ will balance the singularity of the long-range dependent spectrum.

2.2 Measured WAN Traces Now and Then

In [1], Abry and Veitch use their wavelet-based technique to validate previously obtained results showing that measured Ethernet LAN traffic is consistent with long-range dependence and to estimate the corresponding values of the Hurst parameter. In this paper, we use their methodology primarily for investigating the scaling behavior of measured WAN traffic traces; that is, for identifying regions where the scaling property (2) holds for a range of time scales, for detecting changes in scaling behavior, and for finding ranges of time scales with more complex scaling patterns than those captured by Equation (2).

Packet Traces	Year	Time Stamp Accuracy	% WWW (packets)
Bellcore	Jan.1990	1msec	0%
LBL	Jan.1994	1msec	3%
Bellcore	Dec.1994	10μsec	10%
MH (AT&T-Labs)	Feb.1997	1msec	27%
FP (AT&T-Labs)	Aug.1997	1msec	32%
TCP Connection Traces			% WWW (connections)
lbl1	Fall 1993	1msec	0%
lbl2	Fall 1993	1msec	0%
cmu	Jun.1995	1msec	22%
mh (AT&T-Labs)	Feb.1997	1msec	80%
fp (AT&T-Labs)	Aug.1997	1msec	87%

Table 1: Packet-level and TCP connection-level WAN traces used in our study.

The five hour-long packet-level WAN traces used in our study are summarized in the top part of Table 1. The Bellcore'90 trace pre-dates WWW, and TELNET and FTP (as well as NNPT) comprised the main applications at that time. Also note that in 1990, the NSFNET backbone was T1-based; i.e., running at 1.5 Mbps. Both the LBL'94 and the Bellcore'94 WAN traces contain some Web traffic (close to 10% for the Bellcore trace) and represent measurements of Internet traffic as the Internet is going through drastic changes (e.g., emergence of WWW, transition to a 45 Mbps backbone). WWW is the main WAN application and makes up a major portion of the traffic in both the Murray

Hill (MH) AT&T-Labs'97 and Florham Park (FP) AT&T-Labs'97 traces and by that time, the privatization of the Internet is well on its way.

Selecting three representative traces which span seven years of WAN traffic, with 0%, 10%, and 30% or more Web traffic, respectively, Figure 1 (left column) shows the scaling analyses for the time series representing the number of packets per ten milliseconds for the Bellcore'90, Bellcore'94, and FP'97 WAN traces; the corresponding analysis for the number of bytes per 10 milliseconds time series is given in the right column in Figure 1. We have plotted the energy E_j as a function of the resolution level 2^j, on a $\log_2 - \log_2$-scale; i.e., $\log_2 E_j$ as a function of scale j. To clarify the time scales which correspond to scales j, we have included the time (in seconds) on the top axis of each graph. The scaling analysis of a traffic trace which is asymptotically self-similar (or, equivalently, exhibits long-range dependence) will, for large times/scales, result in a linear relationship between $\log_2 E_j$ and the scale j. If the trace is exactly self-similar (i.e., the spectral density characterization of long-range dependence in Equation (1) holds for all frequencies), a plot of $\log_2 E_j$ vs. j will show a linear relationship for all scales. As can be seen, all plots show an approximate linear relationship between $\log_2 E_j$ and scale j for sufficiently large time scales, i.e., the corresponding WAN traces are fully consistent with asymptotic self-similarity or long-range dependence, irrespective of where the traces were collected and whether or not they contained no, some or mostly Web traffic. However, while pre-Web traffic (i.e., Figure 1, top row) seems, in fact, to be consistent with exact self-similarity (note that the linear relationship holds throughout the whole range of observed time scales, from tens of milliseconds to seconds and minutes and beyond), the more recently collected WAN traces exhibit in addition a non-trivial scaling behavior for small time scales (typically below a few hundreds of milliseconds) that is distinctly different from the large-time scaling behavior. Moreover, the "change-point" shows up very clearly as a pronounced "knee" in the graphs in the middle and bottom rows in Figure 1, approximately at scale $j = 5$ corresponding to about 500 milliseconds, irrespective of whether we consider the packet counts or byte counts.

While our finding of the large-time scaling is not new and simply confirms earlier studies by Paxson and Floyd [10] (see also Willinger et al. [14]) who first observed the asymptotically self-similar or long-range dependent nature of measured WAN traffic, our empirical observation of the existence of a pronounced small-time scaling behavior that's clearly different from the large-time scaling phenomenon is new and appears to have gone unnoticed in earlier empirical studies of measured WAN traffic that relied exclusively on techniques such as the variance-time analysis and/or the R/S method. Subsequently, the results of our wavelet-based scaling analysis of a historical collection of WAN traces imply that to fully understand and accurately describe modern WAN traffic, exactly self-similar processes are not sufficient and need to be augmented or replaced by traffic models that allow for a richer scaling behavior.

3 Physical Explanations for Scaling in WAN Traffic

3.1 Large-Time Scaling Phenomenon: Application Layer

In their WAN traffic study, Paxson and Floyd [10] also looked into plausible physical explanations for the observed asymptotically self-similar nature (i.e., large-time scaling phenomenon) of measured WAN traffic. Focusing on the applications, they report on an extensive analysis of FTP and TELNET traffic traces which shows Poisson session arrivals and heavy-tailed session durations or sizes with infinite variance. In view of these

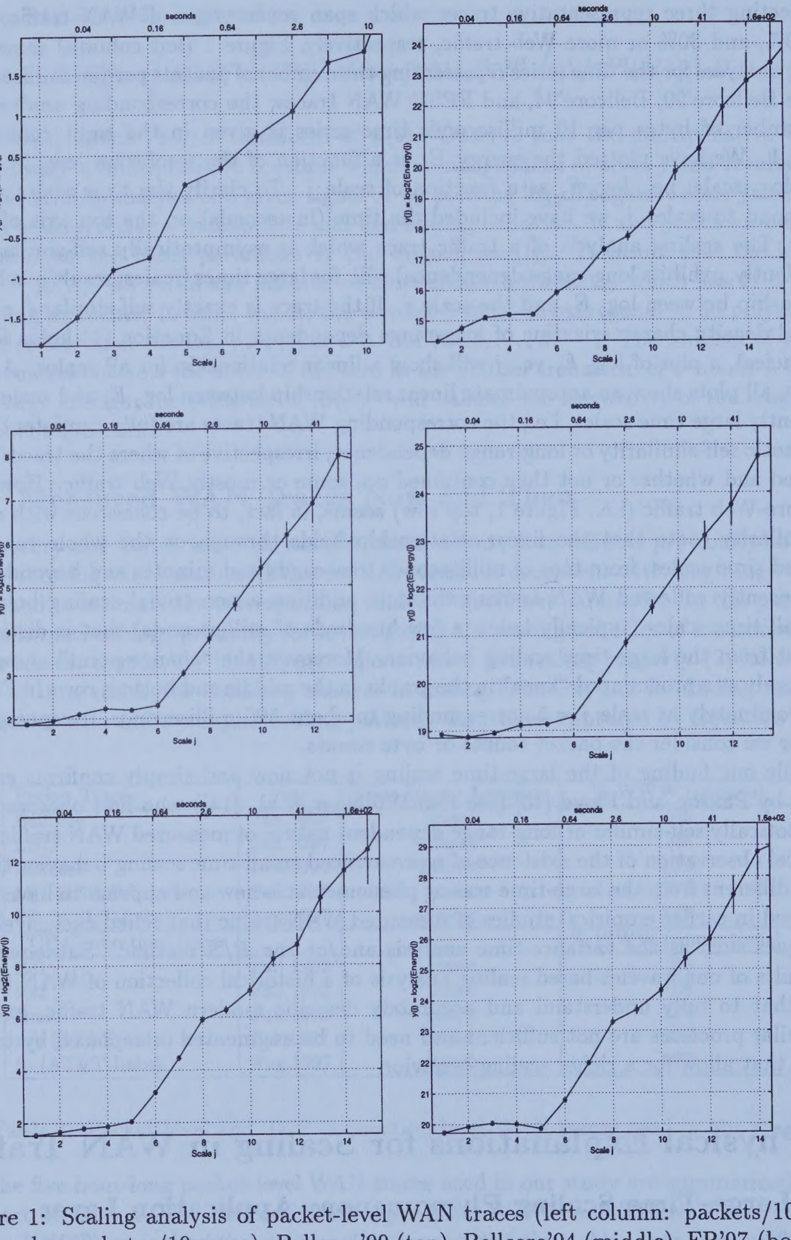


Figure 1: Scaling analysis of packet-level WAN traces (left column: packets/10 msec; right column: bytes/10 msec): Bellcore'90 (top), Bellcore'94 (middle), FP'97 (bottom).

empirically observed application-layer traffic characteristics, a construction due to Cox [3], also known as an *immigration death process* or *M/G/∞ queueing model*, provides a structural modeling approach for pre-Web WAN traffic that is (i) mathematically rigorous, (ii) consistent with measured packet-level WAN traffic at the macroscopic (i.e., aggregated over all individual connections) as well as microscopic (i.e., for individual connections) level, and (iii) highlights the intimate connection between high variability (i.e., heavy-tailed phenomenon) at the microscopic level and (asymptotic) self-similarity (i.e., large-time scaling phenomenon) at the macroscopic level. For details, see for example [13]. In a longer version of this paper [6], we use a plausible, measurable substitute for Web session information, and show that Cox's construction remains valid as a structural modeling approach for WAN traffic, even when the latter is dominated by the Web – and as long as we are only concerned with application and packet level information and ignore other layers in the networking hierarchy, for example TCP layer or IP layer (see Section 3.2 below).

3.2 Large-Time Scaling Phenomenon: TCP Layer

To date, structural modeling approaches for WAN traffic (e.g., Cox's construction) have focused almost exclusively on explaining the observed large-time scaling phenomenon in packet-level WAN traffic in terms of the phenomenon of high variability or infinite variance at the application level. These approaches reflect a source-centered view of traffic modeling, where sources are assumed to simply inject (at a constant rate) as many packets into the network as is required by a user session, and they essentially ignore the impact that the applications (e.g., downloading several Web pages) and networks (e.g., TCP dynamics) have on the flow of packets that individual sources attempt to transmit over the network. In this subsection, we focus exclusively on the transport layer and study the characteristics of measured WAN traffic at the TCP connection level; that is, we investigate the statistical properties of the process representing the number of TCP connection arrivals per time unit. We illustrate our analysis of WAN traffic at the TCP connection level with four hour-long traces whose pertinent statistics are summarized in the lower part of Table 1. Note that lbl2'93 contains an hour worth of all TCP connections from LBL 7 in [10], while lbl1'93 consists of only those TCP connections in lbl1'93 that correspond to successful TELNET or FTP control sessions. As in Section 2.2, we can again study the changing nature of WAN traffic, this time at the level of TCP connection arrivals, over a period where WANs have undergone drastic changes.

Performing the same wavelet-based scaling analysis as in Section 3.1, Figure 2 shows the scaling properties for the time series representing the number of TCP connection arrivals per second, for all five connection-level WAN traces. Starting with the top left plot in Figure 2, we observe a trivial (i.e., a value of the Hurst parameter of about 0.5) scaling behavior for the lbl1'93 data set across all scales. However, when moving from the lbl1'93 to the lbl2'93 data set, i.e., when considering all TCP-connections arrivals (including those spawned by FTP control connections and SMTP or email), we see in Figure 2 (top right plot) a drastic change, namely the appearance of a non-trivial scaling region for the large time scales. The difference between the two plots in the top row of Figure 2 can be attributed to the facts that SMTP-triggered TCP connections dominate and that FTP-control connections typically spawn a number of TCP connections within the FTP session. Moving on to the cmu'1995 trace and analyzing only the HTTP-related portion of the TCP connections, Figure 2 (middle row) shows yet another change in the scaling behavior; Web traffic at the level of TCP connection arrivals no longer exhibits trivial scaling at the small scales but shows across all scales the same non-trivial scaling

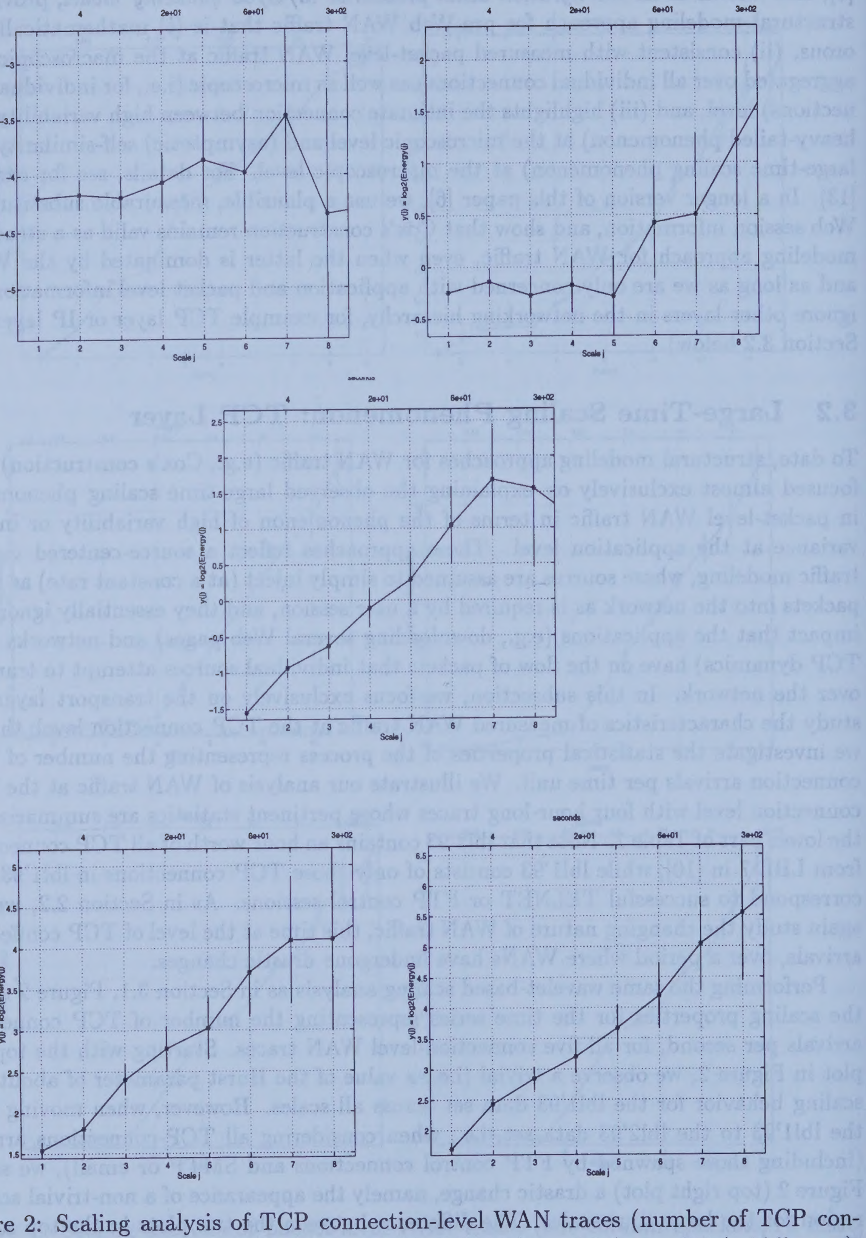


Figure 2: Scaling analysis of TCP connection-level WAN traces (number of TCP connection arrivals per second): lbl1'93 (top left); lbl2'93 (top right); cmu'95 (middle row); mh'97 (bottom left), fp'97 (bottom right).

that we have seen in the pre-Web lbl2'93 trace for large time scales; that is, the traffic representing the number of HTTP-related TCP connection arrivals per second is (exactly) self-similar for time scales of one second and beyond. As can be seen from the remaining two plots in Figure 2 (i.e., bottom row), this observation is confirmed and becomes even more pronounced as we move to the latest connection-level WAN traces and perform the same scaling analysis for the HTTP-related portion of TCP connections. These results confirm earlier findings by Feldmann [5] and shows in addition the gradual appearance over the last few years of non-trivial scaling in WAN traffic at the transport layer, from trivial scaling to non-trivial large-time scaling to non-trivial scaling for all scales. It will be interesting to see, how this empirically observed scaling phenomenon at the transport layer will be affected as a growing portion of Internet traffic will be "multicast", where a single sender transmits to multiple receivers, as other new transport protocols come along, or as the current characteristics of a "typical" Web-page change.

As far as structural modeling approaches for WAN traffic is concerned, the observed self-similar nature of WAN traffic at the TCP connection level clearly demonstrates the limitations of constructions a la Cox that allow for no other "within-session" structure than sending packets at constant rate. In particular, Cox-like constructions are unable to account for the observed TCP connection level characteristics and are, therefore, inadequate to capture small-time scaling phenomena as observed in Section 2.2. However, recent work by Kurtz [8] provides new insights into the problem of allowing for a more flexible "within-session" structure. His construction explicitly allows for the fragmentation of an application-layer session into one or more transport-layer connections (e.g., an Web-session being segmented into a number of TCP connection, interspersed with "idle periods" representing periods of user inactivity or "think time"; see [2]). From the perspective of structural modeling approaches for WAN traffic, Kurtz's construction is an improvement over Cox's construction because it can accommodate a number of relatively complex "within-session" structures. However, at this time, it remains unclear how one would account for the observed self-similar nature of the TCP connection arrival counts or how such a structure (if it exists) would impact the limiting result.

3.3 Small-Time Scaling Properties: Preliminary Observations

We have seen in Section 2.2 that modern-day packet-level WAN traffic exhibits small-time scaling features that are distinctly different from the observed large-time scaling phenomenon. Phenomenological or structural modeling approaches for WAN traffic, such as Cox's or Kurtz's constructions, successfully explain the latter phenomenon in terms of the observed high variability of WAN traffic at the application level, but have failed so far to capture the former. As an illustration of the insights gained from measuring and analyzing high-time resolution WAN packet traces and the ensuing difficulties in trying to understand and describe in a compact manner the small-time scaling behavior in packet-level WAN traffic, we show in Figure 3 the results of a wavelet-based scaling analysis for a WAN trace collected in July'97 from a working commercial network. For this WAN traffic trace, the guaranteed time stamp accuracy is a few micro-seconds, and allows for a scaling analysis down to the millisecond scale. Somewhat surprisingly, the plot shows yet further scaling structure on the very fine time scales, distinctly different from the observed scaling properties over intermediate time scales (see Figure 1), and still different from the well understood large-time scaling feature. This observation should serve as a reminder that even though "black box" models such as fractional ARIMA(p, d, q) processes can often be tweaked so as to reproduce a certain prescribed scaling behavior, they do so without providing any insights into the origins of the fitted scaling behavior. Moreover,

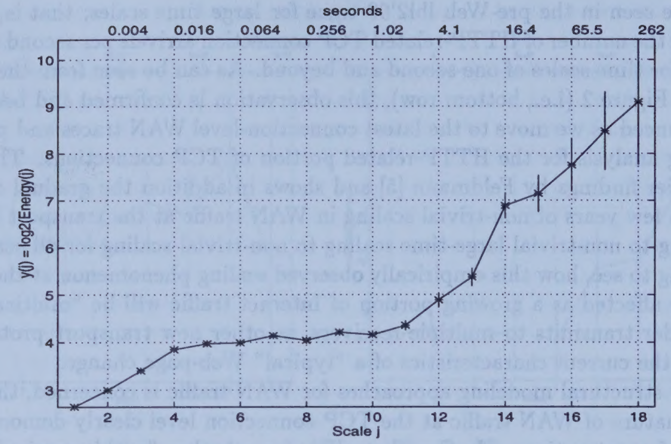


Figure 3: Scaling analysis for a high-time resolution WAN trace (time stamp accuracy: of a few microseconds).

if a more detailed scaling analysis reveals yet further scaling features (as in the case in Figure 3), these "operational" modeling approaches simply result in replacing one black box by another.

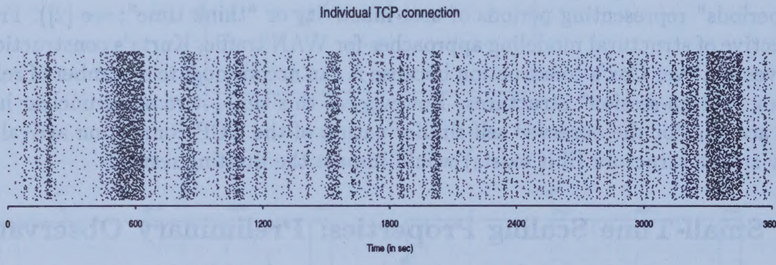


Figure 4: Textured plot for an individual TCP connection. The connection lasted for an entire hour, and consists of 14,305 packets.

From a networking perspective, the presence of non-trivial scaling behavior at small time scales comes as no surprise and is intimately related to the intricacies of the different protocols that rule the flow of traffic at each layer in the networking hierarchy. These protocols interact in non-trivial ways with one another, and are responsible for creating complex interactions between the network on the one side and the sources on the other side. Figure 3.3 illustrates the richness in structure that exists within a single TCP connection via a textured plot and brings out very clearly the high variability of "within-TCP-connection" activity levels and behaviors. The challenge for structural modeling approaches for WAN traffic is to (i) understand the underlying mechanisms that give rise to such rich structures at the transport layer and possibly at the other layers in the networking hierarchy as well, (ii) describe the relevant aspects of the observed structure in a compact and parsimonious manner, and (iii) provide mathematical constructions that reflect the underlying mechanisms and, at he same time, continue to capture the well-understood large-time scaling behavior.

4 Conclusions and Outlook

Currently considered structural modeling approaches for WAN traffic, especially two constructions due to Cox and Kurtz, respectively, are well-suited to explain the large-time scaling features in terms of the well-understood application layer traffic characteristic. However, in their current form, neither of these approaches are capable of explaining the small-time scaling properties, nor are they able to reproduce the self-similar nature of traffic at the TCP layer in an intrinsic and natural manner. The most promising of the current approaches appears to be Kurtz's construction, mainly because in its general form, it allows for the possibility of very general "within-session" packet arrival patterns, and because to date, the limiting regime (i.e., aggregation over many sessions) is only known for a very special cases, all of which can only account for large-time scaling properties. The challenge is to incorporate a more "network-centered" view of traffic modeling into Kurtz's construction (as opposed to further pursuing the traditional "source-centered" approach; see Section 3.2), so as to better account for the impact that the network has on the sources.

One plausible way of making Kurtz's construction more network-oriented – without getting into the details of how the different protocol at the different layers in the networking hierarchy work and interact with one another – is to view WANs or other networks, together with their protocols and controls, as defining the deterministic mechanisms and rules of a process that fragments units of information at one layer into smaller units at the next layer etc. That is, networks act as *cascades* by fragmenting, for example, a Web session into a number of TCP connections (determined by the HTTP protocol) which, in turn, are further fragmented into several HTTP requests (the mechanism for this is provided by the HTTP protocol) which, in turn, are fragmented into IP flows, and so on. Note that during this fragmentation process, the total number of information units (in our example, this would be the total number of bytes transmitted during the Web-session) remains essentially preserved and hence, the cascade analogy is appropriate. Cascades are an important paradigm in the theory of *multifractals* as they can account for many multifractal or self-similar measures, that is, measures with a nontrivial multifractal structure (for an overview, definitions and properties of multifractals, see for example Gupta and Waymire [7] and Evertsz and Mandelbrot [4]). Although the multifractal approach to understanding and describing network traffic appears to be a radical departure from the more conventional structural traffic modeling approaches pursued to date, there exists a close link between Kurtz's construction and random cascade models: simply allow the "within-session" traffic patterns (for details, consult [8]) to be multifractal measures, generated via appropriately defined random cascade models. In addition, there already exists empirical evidence in favor of non-trivial multifractal properties of measured WAN TCP traffic – see the recent work by Riedi and Levy-Vehel [11] and the discussion in [12]. Although multifractals and random cascade models are new traffic modeling paradigms, the motivation for studying them in the network context comes directly from the desire for physically meaningful descriptions of empirically observed phenomena in measured network traffic. In view of the already existing empirical evidence, these new modeling paradigms are more then speculation and offer unique opportunities for studying challenging new mathematical and statistical problems. At the same time they point towards new ways of understanding and describing real-world network traffic, resulting in new engineering insights that can be expected to be of practical relevance for a wide range of network engineering tasks.

Acknowledgments

Many of the traces considered in this paper were collected using the *tcpdump* packet capture tool developed by V. Jacobbson, C. Leres and S. McCanne and available via anonymous ftp to `ftp.ee.lbl.gov`. To extract TCP connection information from the traces, we relied on V. Jacobson's *tcp-conn* and/or V. Paxson's *tcp-reduce* tools, both of which are available from `http://ita.ee.lbl.gov/index.html`.

The Bellcore traces were collected by D. V. Wilson, the LBL/lbl traces were gathered by V. Paxson and are available from `http://ita.ee.lbl.gov/index.html`, and we thank S. Alexander and S. Gao from AT&T Labs-Research for making the MH and FP traffic collection possible. We also acknowledge the help of many of our colleagues at AT&T Labs, especially of J. Friedmanns, with the data collection effort.

References

[1] P. Abry and D. Veitch. Wavelet analysis of long-range dependent traffic. To appear in IEEE Transactions on Information Theory.

[2] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic - evidence and possible causes. *Proceedings of Sigmetrics'96*, pp. 160–169, 1996.

[3] D. R. Cox. Long-range dependence: A review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, pages 55–74. Iowa State University Press, Ames, Iowa, 1984.

[4] C. J. G. Evertsz and B. B. Mandelbrot. Multifractal measures. In H.-O. Peitgen, H. Jurgens and D. Saupe, editors, *Chaos and Fractals: New Frontiers in Science*, Springer-Verlag, New York, 1992.

[5] A. Feldmann. Modelling characteristics of TCP connections. Preprint, 1996.

[6] A. Feldmann, A. C. Gilbert, W. Willinger and T. G. Kurtz. The changing nature of network traffic: Scaling phenomena. Preprint, 1997.

[7] V. K. Gupta and E. C. Waymire. A statistical analysis of mesoscale rainfall as a random cascade. *Journal of Applied Meteorology* **32**, pp. 251–267, 1993.

[8] T. G. Kurtz. Limit theorems for workload input models. In F. P. Kelly, S. Zachary, and I. Ziedins, editors, *Stochastic Networks: Theory and Applications.* Clarendon Press, Oxford, 1996.

[9] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson. On the self-similar nature of Ethernet traffic (Extended Version). *IEEE/ACM Transactions on Networking* **2**, pp. 1–15, 1994.

[10] V. Paxson and S. Floyd. Wide area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking* **3**, pp. 226–244, 1995.

[11] R. H. Riedi and J. Levy Vehel. Multifractal properties of TCP traffic: A numerical study. Preprint, 1997.

[12] M. S. Taqqu, V. Teverovsky and W. Willinger. Is network traffic self-similar or multifractal? *Fractals* **5**, pp. 63-73, 1997.

[13] W. Willinger, V. Paxson, and M. S. Taqqu. Self-similarity and heavy tails: Structural modeling of network Traffic. To appear in R. Adler, R. Feldman, and M. S. Taqqu, editors, *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions*, Birkhauser Verlag, Boston.

[14] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking* **5**, pp. 71–86, 1997.

Multiuser Demodulation: Effective Interference, Effective Bandwidth and Capacity

David Tse, U.C. Berkeley, dtse@eecs.berkeley.edu
Stephen Hanly, U. of Melbourne, s.hanly@ee.mu.oz.au

Abstract

While much of the research in multiuser receivers is performed at the physical layer, the capacity of networks with multi-user receivers and the associated resource allocation problems are less well-understood. In this paper, we show that under some conditions, the capacity of a single cell for several important receivers can be very simply characterized via a notion of *effective bandwidth*: the QoS requirements of all the users can be met if and only if the sum of the effective bandwidths of the users is less than the total number of degrees of freedom in the system. The number of degrees of freedom is the processing gain in a spread-spectrum system and the number of antenna elements in an antenna array. The effective bandwidth of a user depends only on its own QoS requirement, expressed in terms of the desired signal-to-interference ratio, and has a very simple formula for each of the various receivers that we consider.

1 Introduction

To meet the challenges of wireless communications, there have been intense efforts in developing more sophisticated physical layer communication techniques. A significant thrust of work has been on developing *multi-user* receiver structures which mitigate the interference between users in spread spectrum systems. (See for example [12, 5, 6, 14, 7, 9, 10].) Unlike the conventional matched filter receiver used in the IS-95 CDMA system, these techniques take into account the structure of the interference from other users when decoding a user. Another important line of work is the development of processing techniques in systems with antenna arrays. While spread-spectrum techniques provide *frequency diversity* to the wireless system, antenna arrays provide *spatial diversity*, both of which are essentially *degrees of freedom* through which communication can take place.

Despite significant work done in the area, there is still much debate about the network capacity of the various approaches to deal with multi-user interference in spread-spectrum and multi-antenna systems. One important reason is that the networking level problems of resource allocation and power control are less well understood in the context of multi-user techniques than with more traditional multi-access schemes, such as TDMA, FDMA and conventional CDMA systems. For example, in a TDMA or FDMA system, the network resource is shared among users via disjoint frequency and time slots, and this provides a simple abstraction for resource allocation problems at the networking layer. Such clean separation between the networking and the physical layers does not exist when more sophisticated multi-user techniques are used. This in turn hampers the understanding of the capacity of networks with multi-user receivers and of the associated network-level resource allocation problems such as call admissions control, cell handoffs and resource allocation for bursty traffic.

In this paper, we will show that under some conditions, a simple abstraction of the amount of resource consumed by a user is indeed possible for several important multi-user receivers. The specific scenario is a set of power-controlled mobile users communicating to a base-station in a single cell. Assuming that each user's QoS can be expressed in terms of a target signal-to-interference ratio (SIR), we will show that a notion of *effective bandwidth* can be defined such that the QoS requirements of all the users can be met if and only if the sum of the effective bandwidths of the users is less than the total number of degree of freedoms in the system. These degrees of freedom can be provided by the processing gain in a spread-spectrum system or the number of antenna elements in a system with an antenna array. These capacity characterizations are simple in that the effective bandwidth of a user depends only on its SIR requirement and nothing else. The effective bandwidth characterizations also provides a succinct basis for performance comparison between different receiver structures.

The outline of the paper is as follows. In Section 2, we will introduce the basic model of a multi-access spread-spectrum system and the structure of the MMSE receiver. In Section 3, we will present our key result, that in a large system with each user using a random spreading sequence, the limiting interference effects under the MMSE receiver can be calculated as if they were additive; to each interferer can be ascribed a level of *effective interference* that it provides to the user to be decoded. In section 4, we apply this result to study the performance under power control and obtain a notion of *effective bandwidth.* In Section 5, we obtain analogous results for the decorrelating receiver. Section 6 contains our conclusions. Analogous results can be obtained for systems with antenna diversity, but are left out in this conference version due to space limitations.

2 Basic Spread-Spectrum Model and the MMSE Receiver

In a spread-spectrum system, each of the user's information or coded symbols is spread onto a much larger bandwidth via modulation by its own *signature* or *spreading sequence.* The following is a model for a symbol-synchronous multi-access spread-spectrum system:

$$\mathbf{Y} = \sum_{m=1}^{M} X_m \mathbf{s}_m + \mathbf{W},$$

where $X_m \in \Re$ and $\mathbf{s}_m \in \Re^L$ are the transmitted symbol and signature spreading sequence of user m respectively, and $\mathbf{W}$ is $N(0, \sigma^2 I)$ background Gaussian noise. The length of the signature sequences is L, which one can also think of as the number of degrees of freedom or diversity. The received vector is $\mathbf{Y} \in \Re^L$. We assume the X_m's are i.i.d. and that $E[X_m] = 0$ and $E[X_m^2] = P_m$, where P_m is the received power of user m.

Rather than looking at *symbol-by-symbol* detection, we are interested in the more general problem of *demodulation*, extracting good estimates of the (coded) symbols of each user as soft decisions to be used by the channel decoder [10]. From this point of view, the relevant performance measure is the signal-to-interference ratio (SIR) of the estimates.

We shall now focus on the demodulation of user 1, assuming that the receiver has already acquired the knowledge of the spreading sequences. The optimal demodulator that generates a soft decision $\hat{X}_1$ maximizing the signal-to-interference ratio (SIR):

$$\beta_1 = \frac{E[X_1^2]}{E[(X_1 - \hat{X}_1)^2]}$$

is the MMSE receiver [7, 9, 10].

As a comparison, note that the conventional CDMA approach simply matches the received vector to $\mathbf{s}_1$, the signature sequence of user 1. This is indeed the optimal receiver when the interference from other users is white. However, in general the multi-access interference is not white and has structure as defined by $\mathbf{s}_2, \mathbf{s}_3, \ldots, \mathbf{s}_M$, assumed to be known to the receiver. The MMSE receiver exploits the structure in this interference in maximizing the SIR of user 1.

The formulae for the MMSE demodulator and its performance are well known:

$$X_{mmse}(\mathbf{Y}) = \frac{1}{\mathbf{s}_1^t(SDS^t + \sigma^2 I)^{-1}\mathbf{s}_1}\mathbf{s}_1^t(SDS^t + \sigma^2 I)^{-1}\mathbf{Y} \tag{1}$$

and the signal to interference ratio for user 1 is

$$\beta_1 = \mathbf{s}_1^t(SDS^t + \sigma^2 I)^{-1}\mathbf{s}_1 P_1 \tag{2}$$

where $S := [\mathbf{s}_2, \ldots, \mathbf{s}_M]$ and $D := \text{diag}(P_2, \ldots, P_M)$.

3 Performance Under Random Spreading Sequences

Eqn. (2) is a formula for the performance of the MMSE receiver, which one can compute for specific choice of signature sequences. However, it is not easy to obtain qualitative insights directly from the formula. For example, the effect of an individual interferer on the SIR for user 1 cannot be seen directly from this formula. In practice, it is often reasonable to assume that the spreading sequences are *randomly* and independently chosen (see eg. [8]). For example, they may be pseudorandom sequences, or the users choose their sequences from a large set of available sequences as they are admitted into the network. In this case, the performance of the optimal demodulator can be modeled as a random variable, since it is a function of the spreading sequences. In this section, we will show that, unlike the deterministic case, there is a great deal of analytical information one can obtain about this random performance in a large network. In the development below, we will assume that though the sequences are randomly chosen, they are known to the receiver once they are picked. In practice, this means that the change in the spreading sequences is at a much slower time-scale than the symbol rate so that the receiver has the time to acquire the sequences. (There are known adaptive algorithms for which this can even be done blindly; see [3].) However, the *performance* of the MMSE receiver depends on the initial choice of the sequences and hence random.

The model for random sequences: let $\mathbf{s}_m = \frac{1}{\sqrt{L}}(V_{1m}, \ldots, V_{Lm})^t$, m=1, ...M. The random variables V_{km}'s are i.i.d., zero mean and variances 1. The normalization by $\frac{1}{\sqrt{L}}$ ensures that $E[||\mathbf{s}_m||^2] = 1$. In practice, it is common that the entries of the spreading sequences are 1 or -1, but we want to keep the model general so that we can later apply our results to problems with other modes of diversity.

Our results are asymptotic in nature, for a large network, i.e. $M \to \infty$. Clearly, the SIR will go to zero if the spreading length is fixed. Therefore, it is reasonable to scale up L as well, keeping the number of users per degree of freedom (equivalently, per unit bandwidth), $\alpha \equiv \frac{M}{L}$, fixed. We also assume that as we scale up the system, the empirical distribution of the powers of the users converge to a fixed distribution, say $F(P)$. The following is our main result, giving the asymptotic information about the SIR for user 1.

Theorem 3.1 *Let $\beta_1^{(L)}$ be the (random) SIR of the MMSE receiver for user 1 when the spreading length is L. Then $\beta_1^{(L)}$ converges to β_1^* in probability as $L \to \infty$, where β_1^* is the unique solution to the equation:*

$$\beta_1^* = \frac{P_1}{\sigma^2 + \alpha \int_0^\infty I(P, P_1, \beta_1^*) dF(P)} \tag{3}$$

and $I(P, P_1, \beta_1^*) \equiv \frac{PP_1}{P_1 + P\beta_1^*}$

Heuristically, this means that in a large system, the SIR β_1 is deterministic and approximately satisfies:

$$\beta_1 \approx \frac{P_1}{\sigma^2 + \frac{1}{L}\sum_{i=2}^{M} I(P_i, P_1, \beta_1)} \tag{4}$$

where as before P_i is the received power of user i. This result yields an interesting interpretation of the effect of each of the interfering user on the SIR of user 1: for a large system, the total interference can be decoupled into a sum of the background noise and an interference term from each of the other user. (The factor $\frac{1}{L}$ results from the processing gain of user 1.) The interference term depends only on the received power of the interfering user, the received power of user 1 and the attained SIR. It does not depend on the other interfering users except through the attained SIR β_1.

To gain more insights into this concept of effective interference, it is helpful to compare the situation with that when the conventional matched filter $\mathbf{s}_1$ is used for the demodulation. For that case, we have the following proposition, in parallel with Theorem 3.1:

Proposition 3.2 *Let $\beta_{1,MF}^{(L)}$ be the (random) SIR of the conventional matched filter receiver for user 1 when the spreading length is L. Then as $L, M \to \infty$ with $\frac{M}{L} \to \alpha$, $\beta_{1,MF}^{(L)}$ converges almost surely to*

$$\beta_{1,MF}^* = \frac{P_1}{\sigma^2 + \alpha \int_0^\infty P dF(P)}$$

where as before F is the limiting distribution of the powers of the users.

Hence, for large L, the performance of the matched receiver is approximately:

$$\beta_{1,MF} \approx \frac{P_1}{\sigma^2 + \frac{1}{L}\sum_{i=2}^{M} P_i} \tag{5}$$

Comparing this expression with eqn. (4), we see that the interference due to user i is simply P_i in place of $I(P_i, P_1, \beta_1)$. Since the matched receiver filter is independent of the signature sequences of the other users, it is not surprising that the interference is linear in the received powers of the interferers. In the case of MMSE receiver, the filter does depend on the signature sequences of the interferers, thus resulting in the interference being a non-linear function of the received power of the interferer. Also, observe that $I(P_i, P_1, \beta_1) < P_i$, which is expected since the MMSE receiver maximizes the SIR among all linear receivers. But more importantly, we see that while for the conventional receiver, the interference grows unbounded as the received power of the interferer increases, we see that for the MMSE receiver, the effective interference from user i is bounded and approaches $\frac{P_1}{\beta_1}$ as P_i goes to infinity. Thus, while the SIR of the matched filter receiver goes to zero for large interferers' powers, the SIR of the MMSE receiver does not. This is the well-known *near-far resistance* property of the MMSE receiver [7]. The intuition is that as the power of an interferer grows to infinity, the MMSE receiver will null out its signal. While the near-far resistance property has been reported by previous authors, Theorem 3.1 goes beyond that as it not only quantifies the worst-case performance (i.e. large interferer's power) but also the performance for all finite values of the interference. This is useful for example in situations when power control is exercised, as we will now turn to.

4 Capacity and Performance under Power Control

In the present section we consider the user capacity of systems employing multiuser receivers in conjunction with optimal power control. First, we look at the case in which all users require an SIR of exactly β^*, given a processing gain of L degrees of freedom per symbol. For a given number of users we compute the minimum power consumption required to achieve β^* for all users, and then look at the maximum number of users per degree of freedom supportable for a given power constraint under power control. Of particular interest is the maximum number without consideration of power constraints, which we define to be the *capacity* of the system (in terms of number of users per degree of freedom). This coincides with the definition of capacity taken in [1]; "capacity" is then the point at which saturation occurs as we put in so many users that we drive the required power level to infinity. We will show that this capacity is different but finite for both the conventional and the MMSE receivers, thus both are interference-limited systems. As before, our results are asymptotic as the the processing gain L goes to infinity.

Let us focus first on the conventional receiver. Under the matched filter, Prop. 3.2 tells us that asymptotically, users receive the same level of interference, and hence must be received at the same power level to get the same SIR β^*. It is easy to compute that with a processing gain of L and $L\alpha$ users, the common received power required for the conventional receiver asymptotically as $L \to \infty$ is given by

$$P_{mf}(\beta^*) = \frac{\beta^* \sigma^2}{1 - \alpha\beta^*} \tag{6}$$

For a given constraint P on the received power, the maximum number of users supportable is then:

$$\frac{1}{\beta^*} - \frac{\sigma^2}{P} \quad \text{users/degree of freedom}$$

The capacity of the conventional receiver when $P = \infty$ is then

$$C_{mf}(\beta^*) = \frac{1}{\beta^*} \quad \text{users/degree of freedom} \tag{7}$$

(see [1], [2]). Put it another way, as $\alpha \to \frac{1}{\beta^*}$, the system saturates and the required power level goes to infinity.

Now let us turn to the MMSE receiver. To satisfy given target SIR requirements for each user, [4, 11] showed that there is an optimal solution for which the received power of *every user* is minimized; moreover, they gave an iterative algorithm to compute it. However, here we can give an explicit solution and characterize the resulting system capacity.

To begin, we fix the number of users per degree of freedom at α. As in the conventional receiver case, it turns out that the system saturates if α is too high, so we first obtain a necessary and sufficient condition for feasibility. The following theorem shows that in the limit of a large number of degrees of freedom, the system is feasible if and only if the SIR can be met with equal received powers for all users.

Theorem 4.1 *If*

$$\alpha \geq \frac{1+\beta^*}{\beta^*}$$

then there is no distribution F of received powers such that the SIR requirements of all users are satisfied, i.e.:

$$\frac{Q}{\sigma^2 + \alpha \int_0^\infty I(P, Q, \beta^*) dF(P)} \geq \beta^* \quad \textit{for all } Q \textit{ in the support of } F \tag{8}$$

On the other hand, if $\alpha < \frac{1+\beta^*}{\beta^*}$, the SIR requirements of all users can be satisfied and the minimum power solution is having the received powers of all users to be

$$P_{mmse}(\beta^*) = \frac{\beta^* \sigma^2}{1 - \alpha \frac{\beta^*}{1+\beta^*}} \tag{9}$$

Hence, the capacity of the system under MMSE receiver is:

$$C_{mmse}(\beta^*) = \frac{1+\beta^*}{\beta^*} \quad \text{users/degree of freedom.} \tag{10}$$

Moreover, for a given received power constraint P, the maximum number of users that can be supported is to assign each user the same received power, and that number is given by:

$$\frac{1+\beta^*}{\beta^*}(1 - \frac{\beta^* \sigma^2}{P}) \quad \text{users/degree of freedom.}$$

Contrasting (6) and (7) with (9) and (10), we note that if α is feasible for both types of receiver, then the MMSE power consumption is less than the matched filter power consumption, and the MMSE has potentially much greater capacity. Indeed, if $\alpha < 1$ then we can take β^* arbitrarily high without saturating the MMSE receiver, whereas the the conventional receiver saturates as $\beta^* \uparrow \frac{1}{\alpha}$. Thus, in this case, the conventional receiver is interference limited, while the MMSE receiver is not. For fixed β^*, we also note that the MMSE saturates at a higher value of α, yielding a capacity of precisely 1 more user per degree of freedom than the conventional receiver. On the other hand, the relative gain of the MMSE is not so large for small values of β^*.

It is straightforward to generalize our results to the case in which we have J classes, with class j users requiring a SIR of β_j. We denote the number of users of class j by $\alpha_j L$, and again consider the limiting regime $L \uparrow \infty$.

The conventional matched filter results generalize very easily to

$$P_{mf}(j) = \frac{\beta_j \sigma^2}{1 - \sum_{j=1}^{J} \alpha_j \beta_j}$$

where $P_{mf}(j)$ denotes the common received power level of all users of class j (see [1], [2]). Thus, the capacity constraint on feasible values of $(\alpha_1, \ldots, \alpha_J)$ is the linear constraint $\sum_{j=1}^{J} \alpha_j \beta_j < 1$. Furthermore, if class j users have a maximum power constraint that $P_{mf}(j) \le \bar{P}_j$, for each j, then the tighter capacity constraint:

$$\sum_{j=1}^{J} \alpha_j \beta_j \le \min_{1 \le i \le J} \left[1 - \frac{\beta_i \sigma^2}{\bar{P}_i} \right]$$

emerges ([1]). It seems very reasonable to call β_j the bandwidth of class j users, in degrees of freedom per class j user. Let us denote this bandwidth by

$$e_{mf}(\beta_j) \equiv \beta_j \quad \text{degrees of freedom per class j user.}$$

We now show that the MMSE filter results generalize in a similar manner. It is clear in this case also that the minimal power solution consists of the same received power for each class; let all users in class j be received at power P_j. Then the power control equations become

$$\frac{P_j}{\sigma^2 + \sum_{i=1}^{J} \alpha_j I(P_i, P_j, \beta_j)} = \beta_j \quad j = 1, 2, \ldots, J \tag{11}$$

where, as in Theorem 3.1, $I(P_i, P_j, \beta_j) \equiv \frac{P_i P_j}{P_j + P_i \beta_j}$. But (11) implies that $\frac{\beta_j}{P_j}$ is a constant, which allows us to simplify (11) down to

$$P_{mmse}(i) = \frac{\beta_i \sigma^2}{1 - \sum_{j=1}^{J} \alpha_j \frac{\beta_j}{1+\beta_j}} \quad \text{i= 1, 2, \ldots, J.} \tag{12}$$

The capacity constraint for the MMSE receiver with J classes is therefore given by

$$\sum_{j=1}^{J} \alpha_j \frac{\beta_j}{1 + \beta_j} < 1 \tag{13}$$

which is linear in $\alpha_1, \ldots, \alpha_J$.

As above, maximum power constraints provide tighter capacity constraints, and in this context we note that (12) implies that

$$\sum_{j=1}^{J} \alpha_j \frac{\beta_j}{1 + \beta_j} = 1 - \frac{\beta_i \sigma^2}{P_{mmse}(i)} \quad i = 1, 2, \ldots, J.$$

Thus if $P_{mmse}(i) \le \bar{P}_i$ is a maximum power constraint on class i, then the linear constraint

$$\sum_{j=1}^{J} \alpha_j \frac{\beta_j}{1 + \beta_j} \le \min_{1 \le i \le J} \left[1 - \frac{\beta_i \sigma^2}{\bar{P}_i} \right] \quad i = 1, 2, \ldots, J$$

defines the restricted capacity region of the system. It seems very reasonable to define the effective bandwidth of class j users to be $e_{mmse}(\beta_j)$ degrees of freedom per user, where

$$e_{mmse}(\beta_j) \equiv \frac{\beta_j}{1+\beta_j}.$$

Linearity in the matched filter case is a straightforward consequence of the fact that powers add. However, our MMSE effective bandwidth results are rather surprising, as the receiver itself depends on the signature sequences and the received powers of the users. Another interesting observation is that no matter how high β is, the MMSE effective bandwidth of a user is upper bounded by unity. We will gain further insight into why this is so in the next section.

5 The Decorrelator

The decorrelator was in fact the first linear "multi-user detector" introduced by Lupas and Verdu [5]. This receiver is known to be optimal in the worst case scenario in which interferers' powers tend to infinity; its near-far resistance is optimal [6]. The following theorem gives the limiting performance of the decorrelator under random sequences in a large system.

Theorem 5.1 *Let $\beta_1^{(L)}$ be the (random) SIR of the decorrelating receiver for user* 1 *when the spreading length is L. Then $\beta_1^{(L)}$ converges to β_1^* in probability as $L \to \infty$, where β_1^* is given by*

$$\beta_1^* = \begin{cases} \frac{P_1(1-\alpha)}{\sigma^2} & \alpha < 1 \\ 0 & \alpha \geq 1 \end{cases}$$

We observe that as $\alpha \to 1$, i.e. the number of users per degree of freedom approach 1, the SIR goes to zero. Geometrically, as the dimensionality of the orthogonal complement to the span of the interference decreases to zero, the length of the projection of the desired signal onto this orthogonal complement tends to zero, and so in the limit the projected signal is lost in the background noise. This is the high price paid for ignoring the background noise. In contrast, the MMSE receiver can support more users than the degree of freedom as it takes both the interference and the background noise into account.

By comparing Theorem 5.1 and Theorem 3.1, it can be seen that the effective interference for an interferer on user 1 under the decorrelator is $\frac{P_1}{\beta_1}$, which does not depend on the power of the interferer. The theorem states that the capacity constraint on the system is $\alpha < 1$, and this can be generalized to $\sum_{j=1}^{J} \alpha_j < 1$ in the multiple class case.

We also observe that if all users require an SIR of β and employ power control then it is sufficient for each user to be received with power at least $\dfrac{\beta\sigma^2}{1-\alpha}$. Thus, for a given received power constraint $\bar{P}$, the maximum number of users with SIR requirement β supportable is $1 - \frac{\beta\sigma^2}{\bar{P}}$. Similarly, for multiple classes of users with SIR requirement β_j and power constraint $\bar{P}_j$ for each class, then the system can support α_j users (per degree of freedom) from each class if

$$\sum_{j=1}^{J} \alpha_j \leq \min_{1\leq j\leq J} \left[1 - \frac{\beta_j\sigma^2}{\bar{P}_j}\right]$$

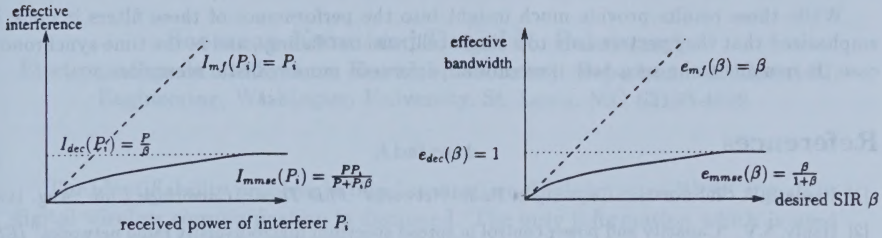


Figure 1: Effective interference and bandwidths for the 3 receivers.

Thus, the capacity region under the decorrelator is given by:

$$\sum_{j=1}^{J} \alpha_j \leq 1 \tag{14}$$

when there are no power constraints, or equivalently, when the background noise power σ^2 goes to zero. Thus, each user occupies an effective bandwidth of 1 degree of freedom, independent of the value of β.

6 Summary of Results

It is illuminating to compare the effective interference and effective bandwidths of the users in the three cases: the conventional matched filter, the MMSE filter, and the decorrelating filter (Fig. 1.). The effective interference under MMSE is non-linear, and depends on the received power P of the user to be demodulated as well as the achieved SIR β. The effective interference under the conventional matched filter is simply P_i, the received power of the interferer. Under the decorrelator, the effective interference is $\frac{P}{\beta}$, independent of the actual power of the interferer. The intuition here is that the decorrelator completely nulls out the interferer, no matter how strong or weak it is.

Assuming perfect power control, we can define effective bandwidths which characterize the amount of network resource a user consumes for a given target SIR. The effective bandwidths under the conventional, MMSE and decorrelating receivers are β, $\frac{\beta}{1+\beta}$ and 1 respectively. We note that the conventional receiver is more efficient than the decorrelator when β is small, and far less efficient when β is large. Intuitively, at high SIR requirements, a user has to transmit at high power, thus causing a lot of interference to other users under the conventional receiver. Not surprisingly, since it is by definition optimal, the MMSE filter is the most efficient in all cases. When β is small, it operates more like the conventional receiver, allowing many users per degree of freedom, but when β is large, each user is decorrelated from the rest, much as in the decorrelator receiver, and therefore the interferers can still occupy no more than 1 degree of freedom per interferer.

The effective bandwidth concept for the MMSE receiver is only valid in the perfectly power-controlled case. By contrast, the concept of effective interference applies with or without perfect power control, and may prove more useful in the multi-cell context.

While these results provide much insight into the performance of these filters it must be emphasized that they pertain only to a single cell, without fading , and in the time-synchronous case. It remains to be seen how these filters perform in more realistic scenarios.

References

[1] Hanly S.V. "Information Capacity of Radio Networks" *PhD Thesis, Cambridge Univ.*, Aug. 1993.

[2] Hanly, S.V. "Capacity and power control in spread spectrum macrodiversity radio networks" *IEEE Trans. on Communications, Vol. 44, No. 2* Feb., 1996.

[3] Honig, M., U. Madhow and S. Verdu, "Blind adaptive multiuser detection", *IEEE Trans. on Information Theory*, July, 1995, pp.944-960.

[4] Kumar, P. and J. Holtzman, "Power control for a spread-spectrum system with multiuser receivers", *Proc. 6th IEEE PIMRC*, Sept. 1995, pp.955-959.

[5] Lupas, R. and S. Verdu, "Linear multiuser detectors for synchronous code-division multiple access", *IEEE Trans. on Information Theory*,IT-35, Jan., 1989, pp.123-136.

[6] Lupas, R. and S. Verdu,"Near-far resistance of multiuser detectors in asynchronous channels," *IEEE Trans. on Communications*, COM-38, Apr. 1990, pp. 496-508.

[7] Madhow, U. and M. Honig, "MMSE interference suppression for direct-sequence spread-spectrum CDMA, *IEEE Trans. on Communications*, Dec., 1994,pp.3178-3188.

[8] Madhow,U. and M. Honig, "MMSE Detection of Direct-Sequence CDMA Signals: Performance Analysis for Random Signature Sequences", submitted to *IEEE Trans. on Information Theory.*

[9] Rapajic, P. and B. Vucetic, "Adaptive receiver structures for asynchronous CDMA systems, *IEEE JSAC*, May, 1994, pp. 685-697.

[10] Rupf, M., F. Tarkoy and J. Massey, "User-separating demodulation for code-division multiple access systems",*IEEE JSAC*, June,1994, pp.786-795.

[11] Ulukus,S. and R.D. Yates, "Adaptive power control and MMSE interference suppression", *Wireless Networks*, accepted for publication.

[12] Verdu,S., "Minimum probability of error for asynchronous Gaussian channels", *IEEE Trans. on Information Theory*, IT-32, Jan.,1986, pp.85-96.

[13] Winters J.H., Salz J., Gitlin R. "The impact of antenna diversity on the capacity of wireless communications systems" *IEEE Trans. on Comm.*, Vol. 42, No. 2/3/4, 1994, pp. 1740-1751.

[14] Xie, Z., R. Short and C. Rushforth, "A family of suboptimum detectors for coherent multi-user communications", IEEE JSAC, May, 1990, pp.683-690.

Identifiability Problem in Blind Separation of Synchronous BPSK Signals in Digital Wireless Communication

Soonman Kwon and Daniel R. Fuhrmann
Electronic Signals and Systems Research Laboratory, Department of Electrical Engineering, Washington University, St. Louis, MO 63130-4899

Abstract

The identifiability problem in blind separation of synchronous BPSK signals in digital wireless communication is discussed. The only information which is used in blind separation in our problem is the finite-alphabet property of the input digital signals. Two kinds of ambiguities in the maximum-likelihood estimate of the signal are analyzed. A new identifiability condition for the solution is introduced and compared to a previously reported condition. We verify the potential applicability of our new identifiability condition by showing that the condition can be achieved with a reasonable number of snapshots.

I. INTRODUCTION

Blind estimation and identification methods in wireless communication have been attracting many researchers in signal processing because of the advances in wireless communication technology and the growth of the telecommunications industry. These techniques are appealing because they eliminate the need for the transmission of training sequences. The early approach to this problem was usually based on a statistical description of the input signals. While techniques based on higher-order statistics were used in the beginning of the research [1], [2], [3], [4], second-order statistics have received more attention recently [5], [6], [7]. Subspace methods [8], [9] based on second-order statistics method have been reported, with promising results.

A more direct approach than the subspace method is a direct blind signal separation method, proposed in several works [10], [11], [12]. These authors just focused on separating the signal directly without determining the channel characteristics. Talwar *et al.*'s work [12] handled this problem by exploiting the finite alphabet (FA) property of digital signals. The identifiability issue was considered first and then two optimization algorithms were proposed for finding the maximum-likelihood estimate of the signal. However, their identifiability condition can only be satisfied with a great amount of snapshots of the transmitted signal. For example, even in the case of 10 incoming BPSK signals, at least 512 columns are needed at a time, for the sampled data matrix at the receiver, to determine the estimate of the transmitted signals. Thus, the data matrix becomes too large to be manipulated in a reasonable time at the receiver.

This paper mainly addresses the identifiability issue in the blind separation problem using the FA property of the input signals. We show the existence of a high-level ambiguity besides the basic ambiguity discussed in [12]. We also derive a tighter identifiability condition for the maximum-likelihood (ML) solution.

This paper will be organized as follows. A mathematical model of the received signal at a sensor array is formulated in Section II. In Section III, the ML estimate of the signal is derived, and the ambiguity issue related to the solution of the estimate is considered. A new identifiability condition for the estimate is proposed and compared with a previously reported condition in view of the probabilities of their outcomes in the same Section. A brief conclusion follows.

II. PROBLEM STATEMENT

We consider d narrowband plane-wave signals which reach an M-sensor array with arbitrary sensor locations through multipath channels. Practically, the reverse-link transmission in mobile wireless communication systems may be considered as one of these cases.

The mathematical model of the system may be formulated as follows. Assuming that the signal from k-th user is represented as $p_k s_k(t)$, and using complex envelope representation, we have the $M \times 1$ received vector signal at the sensor array expressed by

$$\mathbf{z}(t) = \sum_{k=1}^{d} \sum_{l=1}^{L_k} p_k \alpha_{kl} \mathbf{d}(\theta_{kl}) s_k(t - \tau_{kl}) + \mathbf{w}(t) \tag{1}$$

where L_k denotes the number of sub-paths for the k-th signal, α_{kl} and τ_{kl} denote the attenuation and time delay corresponding to l-th sub-path of the k-th signal, respectively, p_k denotes the amplitude of the k-th signal, $\mathbf{d}(\theta_{kl})$ denotes the array response vector to a signal from direction θ_{kl}, $s_k(\cdot)$ denotes the k-th signal waveform, and $\mathbf{w}(\cdot)$ denotes additive white noise vector process with the matrix correlation function $E\{\mathbf{w}(t)\mathbf{w}^{\dagger}(s)\} = \gamma^2 \mathbf{I}\delta(t - s)$ where $\mathbf{I}$ is an identity matrix.

Under the assumption that the bandwidth of the incident signal is much smaller than the reciprocal of the maximum propagation delay, (1) becomes

$$\mathbf{z}(t) = \sum_{k=1}^{d} p_k \mathbf{a}_k s_k(t) + \mathbf{w}(t) \tag{2}$$

where

$$\mathbf{a}_k = \sum_{l=1}^{L_k} \alpha_{kl} e^{-j2\pi f_c \tau_{kl}} \mathbf{d}(\theta_{kl}) \tag{3}$$

and f_c is the carrier frequency.

In digital communication the modulating signal may be given by

$$s_k(t) = \sum_{n=1}^{N} b_k(n) h(t - nT_s) \tag{4}$$

where N is the number of symbols, $\{b_k(\cdot)\}$ is the symbol sequence of the k-th user, T_s is the symbol period, and $h(\cdot)$ is the pulse shape such as a raised cosine waveform.

Assuming that the signals are symbol-synchronous, we have the symbol-rate sampled signal outputs of a matched filter bank for the signal in (2) as

$$\begin{aligned} \mathbf{z}(n) &= \sum_{k=1}^{d} p_k \mathbf{a}_k b_k(n) + \mathbf{w}(n) && (5) \\ &= \mathbf{A}\mathbf{s}(n) + \mathbf{w}(n) && (6) \end{aligned}$$

where $\mathbf{s}(n) = [b_1(n) \cdots b_d(n)]^T$, $\mathbf{w}(\cdot)$ is an additive white noise discrete random vector process with $E\{\mathbf{w}(n)\mathbf{w}^{\dagger}(m)\} = \gamma^2 \mathbf{I}\delta(n - m)$, and $\mathbf{A}$ is an $M \times d$ matrix given by $\mathbf{A} = [p_1\mathbf{a}_1, \cdots, p_d\mathbf{a}_d]$.

Collecting N samples and assuming that the channel is time-invariant over the available N symbol periods, we have

$$\mathbf{Z} = \mathbf{A}\mathbf{S} + \mathbf{W} \tag{7}$$

where $\mathbf{Z} = [\mathbf{z}(1)\cdots\mathbf{z}(N)]$, $\mathbf{S} = [\mathbf{s}(1)\cdots\mathbf{s}(N)]$, and $\mathbf{W} = [\mathbf{w}(1)\cdots\mathbf{w}(N)]$. The problem we consider here is to determine the digital signal matrix $\mathbf{S}$ from the given measurement $\mathbf{Z}$.

Typical array signal processing problems having the structure in (7) usually have the assumption that the column vector spaces of $\mathbf{A}$ lie on the array manifold which is known or can be estimated. Solution techniques for this case have been well developed [13],[14]. However it should be noted that the array response vectors are completely unknown in our model due to the multipath effects. Thus we have to solve this problem in a different way.

Without any useful information about the unknown matrix $\mathbf{A}$, $\mathbf{S}$ should be determined solely by depending on any prior information on $\mathbf{S}$ itself. This is called the *blind separation* problem.

In general, if the digital signals are modeled as unknown deterministic sequences corrupted by spatially and temporally white Gaussian noise, the maximum-likelihood solution is frequently used to estimate $\mathbf{S}$.

III. AMBIGUITIES IN THE SOLUTION FOR THE ML ESTIMATE OF S

The ML estimator for $\mathbf{S}$ may be derived as follows.

Supposing that the digital signals are modeled as unknown deterministic sequences corrupted by spatially and temporally white Gaussian noise, we have the probability density function for the measurement matrix $\mathbf{Z}$ given by

$$f_{\mathbf{Z}}(Z|\mathbf{A},\mathbf{S}) = \pi^{-MN}\gamma^{-2MN}\exp\left\{-\frac{1}{\gamma^2}\sum_{n=1}^{N}\|\mathbf{z}(n)-\mathbf{A}\mathbf{s}(n)\|^2\right\} \tag{8}$$

where $Z = [\mathbf{z}(1)\ \mathbf{z}(2)\ \ldots\ \mathbf{z}(N)]$. Therefore, the ML estimate of the signals is acquired by solving the following least-squares problem:

$$\min_{\mathbf{A},\mathbf{S}\in\mathcal{S}} J \quad \text{where} \quad J = \|\mathbf{Z}-\mathbf{A}\mathbf{S}\|_F^2. \tag{9}$$

The solution $(\hat{\mathbf{A}}, \hat{\mathbf{S}})$ must satisfy

$$\hat{\mathbf{A}} = \mathbf{Z}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T)^{-1} \tag{10}$$

$$\hat{\mathbf{S}} = \arg\max_{\mathbf{S}\in\mathcal{S}} \|\mathbf{Z}\mathbf{P}_S\|_F^2 \tag{11}$$

where $\mathbf{P}_S = \mathbf{S}^T(\mathbf{S}\mathbf{S}^T)^{-1}\mathbf{S}$ and $\|\cdot\|_F$ denotes the Frobenius norm. Note that the function in $\|\cdot\|$ represents the row projection of $\mathbf{Z}$ onto the row subspace of $\mathbf{S}$.

Our goal here is to determine the $\mathbf{S}$ which maximizes the cost function in (11). Since both $\mathbf{A}$ and $\mathbf{S}$ are unknown in our problem, there may be infinite solution sets for $(\mathbf{A}, \mathbf{S})$. Fortunately, we know that the alphabet of the input sequences of the digital wireless communication channel is finite. Define $\mathcal{M}_{p\times q}(r)$ to be the set of all $p \times q$ matrices with elements from the set r. Then we are seeking for the solution pair $(\mathbf{A}, \mathbf{S})$ in the set pair $(\mathcal{M}_{M\times d}(\mathbb{C}), \mathcal{M}_{d\times N}(\mathbb{N}))$ where $\mathbb{C}$ denotes the complex numbers and $\mathbb{N} = \{+1, -1\}$.

A major issue to be considered in this minimization problem is the non-uniqueness of the solution set. Actually, one interesting fact found from the computer simulation is that under the no-noise assumption the number of the norm values of the cost function is much smaller than the number of possible signal matrices $\mathbf{S}$. The number of norm values increases approximately linearly with respect to the increase in d or N whereas the cardinality of $\mathcal{M}_{d\times N}(\mathbb{N})$ is exponentially increasing as d or N increases. Therefore, we investigate the possible ambiguities in the solution of our problem which cause the same norm values.

A. Ambiguities in the solution

The existence of the solution of (7) depends on the structure of the matrices in the equation. In fact, the arbitrariness of the matrix $\mathbf{A}$ causes a question for the identifiability of the solution.

Suppose that there is no noise in the transmission so that the exact solution for (9) becomes the set pair

$$\{(\mathbf{A}, \mathbf{S}) : \mathbf{Z} = \mathbf{AS}\}. \tag{12}$$

Then the problem can be considered as a nonlinear factorization of the observation matrix $\mathbf{Z}$ into factors $\mathbf{A}$ and $\mathbf{S}$. It is well known that this factorization is unique up to permutation if (i) $\mathbf{A}$ and $\mathbf{S}$ are full rank, (ii) the column space of $\mathbf{A}$ lies on the array manifold, and (iii) $d < M$. In our problem it is assumed that $\mathbf{A}$ is full rank. But the condition (ii) is not satisfied since $\mathbf{A}$ is completely unknown. Therefore, this problem can not be solved directly by the well-known numerous methods which basically rely on the direction-finding techniques.

The ambiguities related to our factorization problem are classified into two kinds as follows.

Low-level Ambiguity

Considering the solution of (12), its structure gives an inherent ordering ambiguity with additional sign ambiguity to its solution, since if

$$\mathbf{Z} = \mathbf{AS} \tag{13}$$

then

$$\mathbf{Z} = \mathbf{ATT}^{-1}\mathbf{S} = \bar{\mathbf{A}}\bar{\mathbf{S}} \tag{14}$$

where $\mathbf{T}$ is a $d \times d$ matrix with exactly one nonzero element from $\mathbb{N}$ in each row and column. We call this basic ambiguity a *low-level ambiguity.*

Note that $\mathbf{T}$ may be described as $\mathbf{T} = \mathbf{PD}$ (or $\mathbf{T} = \mathbf{DP}$) where $\mathbf{D}$ is a diagonal matrix with elements from $\mathbb{N}$, and $\mathbf{P}$ is a permutation matrix.

Here we define a term related with this low-level ambiguity for future use.

Definition 1: Two matrices, $\mathbf{S}_i$, $\mathbf{S}_j \in \mathcal{M}_{d\times N}(\mathbb{N})$, are said to be *ambiguity-equivalent* if $\mathbf{S}_i = \mathbf{TS}_j$ and $\mathbf{T} = \mathbf{PD}$. They are also said to be in the same *ambiguity-equivalence class.*

However, there may exist additional ambiguities beyond this low-level ambiguity in our problem because $\mathbf{A}$ is arbitrary. We call it a *high-level ambiguity* here.

High-level Ambiguity

Solving for $\bar{\mathbf{A}}$ in (14), we have

$$\bar{\mathbf{A}} = \mathbf{A}\mathbf{S}\bar{\mathbf{S}}^T(\bar{\mathbf{S}}\bar{\mathbf{S}}^T)^{-1}. \tag{15}$$

Thus

$$\bar{\mathbf{A}}\bar{\mathbf{S}} = \mathbf{A}\mathbf{S}\bar{\mathbf{S}}^T(\bar{\mathbf{S}}\bar{\mathbf{S}}^T)^{-1}\bar{\mathbf{S}} = \mathbf{AS} \tag{16}$$

and

$$\mathbf{A}(\mathbf{S} - \mathbf{S}\bar{\mathbf{S}}^T(\bar{\mathbf{S}}\bar{\mathbf{S}}^T)^{-1}\bar{\mathbf{S}}) = \mathbf{0}. \tag{17}$$

Since $\mathbf{A}$ is full rank, we finally have

$$\mathbf{S} = \mathbf{S}\mathbf{P}_{\bar{S}} \tag{18}$$

where $\mathbf{P}_{\bar{S}} = \bar{\mathbf{S}}^T(\bar{\mathbf{S}}\bar{\mathbf{S}}^T)^{-1}\bar{\mathbf{S}}$. Note that the right side of (18) is the projection of the rows of $\mathbf{S}$ onto the row space of $\bar{\mathbf{S}}$ by the projection matrix $\mathbf{P}_{\bar{S}}$. Therefore, we can write the relation as

$$\mathbf{S} = \mathbf{T}\bar{\mathbf{S}}, \tag{19}$$

where $\mathbf{T}$ is a $d \times d$ matrix.

Rewriting (19) using vector notation, we have

$$\mathbf{S} \equiv \begin{bmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \vdots \\ \mathbf{s}_d^T \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1^T\bar{\mathbf{S}} \\ \mathbf{t}_2^T\bar{\mathbf{S}} \\ \vdots \\ \mathbf{t}_d^T\bar{\mathbf{S}} \end{bmatrix} \equiv \mathbf{T}\bar{\mathbf{S}}. \tag{20}$$

(20) can be rewritten as the transposed form given by

$$\begin{bmatrix} \bar{\mathbf{S}}^T\mathbf{t}_1 & \bar{\mathbf{S}}^T\mathbf{t}_2 & \cdots & \bar{\mathbf{S}}^T\mathbf{t}_d \end{bmatrix} = \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_d \end{bmatrix}. \tag{21}$$

Note that (21) is d sets of simultaneous equations which is solved for the vector set $\{\mathbf{t}_j\}_{j=1}^{d}$. We here assume that $\mathbf{T} \neq \mathbf{PD}$ since it is the case of low-level ambiguity as discussed before. Then what kind of solution may be possible for $\mathbf{T}$ here? A simple case is considered here.

If it is assumed that $\{\mathbf{t}_j\}_{j=1}^{d-1}$ come from the $d-1$ columns of a permutation matrix so that the $d-1$ rows of $\bar{\mathbf{S}}$ are a combination of $d-1$ rows of $\mathbf{S}$, then one set of the simultaneous equations, which is

$$\bar{\mathbf{S}}^T\mathbf{t}_i = \mathbf{s}_i, \tag{22}$$

remains and is just a problem solving a set of simultaneous linear equations. It is well known that if $rank(\bar{\mathbf{S}}^T) = rank([\bar{\mathbf{S}}^T\ \mathbf{s}_i]) = d$, we have unique solution for $\mathbf{t}_i$ which is $\mathbf{t}_i = (\bar{\mathbf{S}}\bar{\mathbf{S}}^T)^{-1}\bar{\mathbf{S}}\mathbf{s}_i$. Note that $\mathbf{s}_i$ can not be equal to any column in $\bar{\mathbf{S}}^T$, since we assumed that $\mathbf{S}$ and $\bar{\mathbf{S}}$ are not in the same ambiguity-equivalence class. Therefore, in the first case, $\mathbf{s}_i$ should be a linear combination of some columns of $\bar{\mathbf{S}}^T$ to guarantee the unique solution for $\mathbf{t}_i$. That results in the fact that $\mathbf{t}_i$ should have at least two non-zero elements inside. Thus it is true that there can exist a high-level ambiguity.

In view of the digital communication technology, the ordering and sign ambiguity (i.e., low-level ambiguity) may not be a serious problem. The sign ambiguity can be resolved by adopting a differential coding scheme. Also the ordering ambiguity is irrelevant because of the association of the n-th column of $\mathbf{A}$ with the n-th row of $\mathbf{S}$. Thus we here define the system $\mathbf{Z} = \mathbf{AS}$ to be *identifiable* if the solution can be solved up to the ambiguity-equivalence classes so that all simultaneous solutions can be written as $\mathbf{AT}$ and $\mathbf{T}^{-1}\mathbf{S}$, with $\mathbf{T}$ of the form $\mathbf{T} = \mathbf{PD}$. We now consider the condition for the identifiability.

B. Identifiability Problem

Talwar *et al.* showed a sufficient identifiability condition rewritten in Theorem 1 [12].

Definition 2: Two vectors $\mathbf{a}_1$, $\mathbf{a}_2$ are said to be *equivalent* (or *distinct*) *up to sign* if $\mathbf{a}_1 = \pm\mathbf{a}_2$ (or if $\mathbf{a}_1 \neq \pm\mathbf{a}_2$).

Theorem 1: Let $\mathbf{Z} = \mathbf{AS}$ where $\mathbf{A}_{M\times d}$ is an arbitrary full-rank matrix with $d \leq M$, and $\mathbf{S}_{d\times N}$ is a full-rank matrix with ± 1 elements. If the columns of $\mathbf{S}$ include all the 2^{d-1} possible distinct (up to a sign) d-vectors with ± 1 elements, then the system $\mathbf{Z} = \mathbf{AS}$ is identifiable.

As we can see in Theorem 1, at least 2^{d-1} columns are needed for the identifiability condition to be satisfied. Thus the size of the observation matrix should be increased exponentially to guarantee the identifiability at some reasonable level as the number of signals d is linearly increased. Actually, if $d = 10$, the the number of samples for the observation matrix should be at least 512. In fact, the actual required number of samples to guarantee the identifiability may become several times of 512. Thus we can easily foresee the computational complexity in this case. Therefore, the investigation for a tighter sufficient condition is necessary to make the condition more applicable.

B.1 An Improved Sufficient Condition for the Identifiability

Some definitions are needed to introduce a new sufficient condition for the identifiability.

Definition 3: A vector set $\mathcal{B}_d$ is defined as the collection of all d-vectors with elements from $\{+1, -1\}$.

Definition 4: A vector set $\mathcal{O}(\mathbf{S})$ is defined as the set of all different column vectors of the matrix $\mathbf{S}$.

Definition 5: A set $\mathcal{H}_d(\mathbf{a}, k) \subset \mathcal{B}_d$ is defined as the collection of all d-vectors which have Hamming distance k from the given d-vector $\mathbf{a}$, and $\mathbf{a}$ itself.

Definition 6: A real matrix $\mathbf{S} \in \mathcal{M}_{d\times N}(\mathbb{N})$ is said to satisfy *Hamming distance condition of k* if, for every $\mathbf{a}_1 \in \mathcal{H}_d(\mathbf{a}, k)$, $\mathbf{a}_1 \in \mathcal{O}(\mathbf{S})$ or $-\mathbf{a}_1 \in \mathcal{O}(\mathbf{S})$.

Theorem 2: If the signal matrix $\mathbf{S}$ satisfies Hamming distance condition of 1, then the system $\mathbf{Z} = \mathbf{AS}$ is identifiable.

Proof: We start with the relation in (21) which may be rewritten as

$$\bar{\mathbf{S}}^T \mathbf{t}_i = \mathbf{s}_i, \quad i \in \{1, 2, \ldots, d\}. \tag{23}$$

Our condition for the identifiability is considered as follows.

Let the reference column vector in $\bar{\mathbf{S}}$ be $[\sigma_d\ \sigma_{d-1}\ \ldots\ \sigma_2\ \sigma_1]^T$. Then the new assumption in Theorem 2 gives a well organized form of (23) for an i which is

$$\left(\begin{array}{c}
\sigma_d t_d + \sigma_{d-1} t_{d-1} + \cdots + \sigma_5 t_5 + \sigma_4 t_4 + \sigma_3 t_3 + \sigma_2 t_2 + \sigma_1 t_1 = b_1 \\
\sigma_d t_d + \sigma_{d-1} t_{d-1} + \cdots + \sigma_5 t_5 + \sigma_4 t_4 + \sigma_3 t_3 + \sigma_2 t_2 - \sigma_1 t_1 = b_2 \\
\sigma_d t_d + \sigma_{d-1} t_{d-1} + \cdots + \sigma_5 t_5 + \sigma_4 t_4 + \sigma_3 t_3 - \sigma_2 t_2 + \sigma_1 t_1 = b_3 \\
\vdots \\
\sigma_d t_d - \sigma_{d-1} t_{d-1} + \cdots + \sigma_5 t_5 + \sigma_4 t_4 + \sigma_3 t_3 + \sigma_2 t_2 + \sigma_1 t_1 = b_d \\
-\sigma_d t_d + \sigma_{d-1} t_{d-1} + \cdots + \sigma_5 t_5 + \sigma_4 t_4 + \sigma_3 t_3 + \sigma_2 t_2 + \sigma_1 t_1 = b_{d+1}
\end{array}\right) \tag{24}$$

where $b_i = \pm 1$, $i = 1, 2, \ldots, d+1$.

Rewrite this set using a notation in Gauss elimination method, and then taking the first equation as the pivot equation in the elimination procedure, we come up to the form given by

$$\left[\begin{array}{ccccccccc}
\sigma_d & \sigma_{d-1} & \sigma_{d-2} & \cdots & \sigma_4 & \sigma_3 & \sigma_2 & \sigma_1 & b_1 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 2\sigma_1 & b_1 - b_2 \\
0 & 0 & 0 & \cdots & 0 & 0 & 2\sigma_2 & 0 & b_1 - b_3 \\
 & & & & \vdots & & & & \\
0 & 2\sigma_{d-1} & 0 & \cdots & 0 & 0 & 0 & 0 & b_1 - b_d \\
2\sigma_d & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & b_1 - b_{d+1}
\end{array}\right]. \tag{25}$$

Note that the terms $\{b_1 - b_i\}_{i=1}^{d}$ can be -2 or 0 if $b_1 = -1$, and 0 or 2 if $b_1 = +1$. Thus from the second row to the last row, we have the solutions such that

$$t_j = \frac{b_1 - b_{j+1}}{2\sigma_j}, \ j = 1, 2, \ldots, d \tag{26}$$

which turn out to be one value among $\{-1, 0, +1\}$ for each j. Now, insert these solutions into the first equation in (25). Then this equation becomes

$$\frac{b_1 - b_{d+1}}{2} + \frac{b_1 - b_d}{2} + \frac{b_1 - b_{d-1}}{2} + \cdots + \frac{b_1 - b_4}{2} + \frac{b_1 - b_3}{2} + \frac{b_1 - b_2}{2} = b_1. \qquad (27)$$

Note that every term in the left side of the equation can be 0 or +1 if $b_1 = +1$, and 0 or -1 if $b_1 = -1$. Thus if one term is not zero, the others should be all zeros. Consequently only one t_j can be nonzero and is either $+1$ or -1. □

In (21), there are actually d sets of these system equations so that we have exactly the same solution for each set. Since we assumed that $\mathbf{S}$ and $\bar{\mathbf{S}}$ are full rank, the columns of $\mathbf{T}$ should be independent from each other to satisfy this assumption. Therefore, $\mathbf{T}$ should have exactly one non-zero element in each row and column to make itself full rank. Therefore our theorem shows a sufficient condition for the identifiability.

Note that the smallest number of columns which fulfills the new condition is $d+1$ in this case.

Now the two identifiability conditions are compared in view of their probabilities of outcomes.

C. Comparison of the Two Conditions for the Identifiability

C.1 The number of different sets which satisfy our identifiability condition

In contrast to the uniqueness of the set which satisfies Talwar's identifiability condition, there are several choices of the sets which satisfy our new identifiability condition as shown in this section.

Definition 7: The set $\mathcal{A}_d$ is defined as the collection of all the d-vectors in $\mathcal{B}_d$ which have +1 as their first elements.

Note that the cardinality of $\mathcal{A}_d$ is just half the cardinality of $\mathcal{B}_d$. Thus, we can let $\mathcal{A}_d \equiv \{\mathbf{a}_1, \cdots, \mathbf{a}_{2^{d-1}}\}$. By introducing this $\mathcal{A}_d$, we can remove the sign ambiguity in the following analysis.

Lemma 1: All of the sets, $\mathcal{H}_d(\mathbf{a}_i, 1) \subset \mathcal{B}_d$, with $\mathbf{a}_i \in \mathcal{A}_d$, $i = 1, 2, \cdots, 2^{d-1}$, are different from each other if $d > 3$.

Proof: Let $\mathcal{H}_d(\mathbf{u}_1, 1)$ and $\mathcal{H}_d(\mathbf{v}_1, 1)$ be two different sets in $\mathcal{B}_d$ and $\mathbf{u}_1$ and $\mathbf{v}_1$ are any two elements of $\mathcal{A}_d$. Then, in order that the two sets have the same elements the Hamming distance between the two reference vectors should be just 1 so that $\mathbf{v}_1 \in \mathcal{H}_d(\mathbf{u}_1, 1)$ and $\mathbf{u}_1 \in \mathcal{H}_d(\mathbf{v}_1, 1)$. If it is true, then there always exists at least one vector which has Hamming distance of 1 with respect to $\mathbf{u}_1$, but Hamming distance of 2 to $\mathbf{v}_1$. If we call this vector $\mathbf{q}$, then $\mathbf{q} \in \mathcal{H}_d(\mathbf{u}_1, 1)$, but $\mathbf{q} \notin \mathcal{H}_d(\mathbf{v}_1, 1)$. □

Corollary 1: In the set $\mathcal{A}_d$, there are 2^{d-1} different combinations of d-vector sets with cardinality $d+1$ which satisfy our identifiability condition.

C.2 Probability issue

Here we consider the calculation of the probabilities of satisfying the identifiability condition for both cases. In the calculation it is assumed that all possible $\mathbf{S}$ are equally likely.

The condition in Theorem 1 (Talwar's condition)

In this case note that we need all 2^{d-1} vectors in $\mathcal{A}_d$, which are $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \cdots, \mathbf{a}_{2^{d-1}}]$, to satisfy the identifiability condition. Therefore the problem becomes an *N-permutation* of a set (call this set Ω) of 2^{d-1} distinct elements each with a repetition number n_i, $i = 1, 2, \cdots, 2^{d-1}$ which satisfy the following condition. If we use $l \equiv 2^{d-1}$ for convenience from here on, then the condition is

$$\sum_{i=1}^{l} n_i = N, \tag{28}$$

where

$$\Omega = \{n_1 \cdot \mathbf{a}_1,\, n_2 \cdot \mathbf{a}_2,\, \cdots,\, n_{l-1} \cdot \mathbf{a}_{l-1},\, n_l \cdot \mathbf{a}_l\},\ n_i \geq 1,\ i = 1, 2, \cdots, l. \tag{29}$$

Here we use the notation, $n \cdot \mathbf{a}$, to represent that there are n of the same $\mathbf{a}$'s in the set.

The number of possible permutations of this case is given by

$$M_1(N, l) = \sum_{i_l=1}^{(N-l+1)} \sum_{i_{l-1}=1}^{(N-l-i_l+2)} \sum_{i_{l-2}=1}^{(N-l-i_l-i_{l-1}+3)} \cdots \sum_{i_2=1}^{(N-\sum_{j=3}^{l} i_j-1)} \frac{N!}{i_l! i_{l-1}! \cdots i_2! \left(N - \sum_{k=2}^{l} i_k\right)!}. \tag{30}$$

Note that the upper bound in each summation and the last multiplicand term in the denominator are just following the constraint given in (28). Now let the event where we have all the elements of $\mathcal{A}_d$ in the N-snapshots be $\mathcal{E}_1$. Then, the probability of the event is exactly given by

$$P_r(\mathcal{E}_1) = \frac{M_1(N, l)}{l^N}. \tag{31}$$

The condition in Theorem 2 (New condition)

The consideration for all the cases of the permutations becomes more complicated here than in the previous case since we have many possible combinations of the sets which satisfy the identifiability condition.

For a fixed set which satisfies the identifiability condition, the number of possible permutations is given by

$$M_2(N, l) = \sum_{i=0}^{N-n} {}_N\mathcal{C}_n \times C_i \times (l - n)^{N-n-2}, \tag{32}$$

where ${}_N\mathcal{C}_n = \frac{N!}{n!(N-n)!}$, $n = d + 1 = \log_2 l + 2$, and

$$C_i = \sum_{i_l=1}^{(i+1)} \sum_{i_{l-1}=1}^{(i-i_l+2)} \sum_{i_{l-2}=1}^{(i-i_l-i_{l-1}+3)} \cdots \sum_{i_2=1}^{(i-\sum_{j=3}^{l} i_j+l-1)} \frac{(n+i)!}{i_l! i_{l-1}! \cdots i_2! \left(i + n - \sum_{k=2}^{l} i_k\right)!}. \tag{33}$$

Again let the event in which we have all the elements we need for the identifiability be $\mathcal{E}_2$. Then we have another loose bound for the probability of $\mathcal{E}_2$ given as

$$P_r(\mathcal{E}_2) > \frac{M_2(N, l)}{l^N}. \tag{34}$$

As claimed in Corollary 1, there exist 2^{d-1} different combinations of the subsets which satisfy the identifiability condition. In the following, we show one example of the exact result obtained by computer calculation.

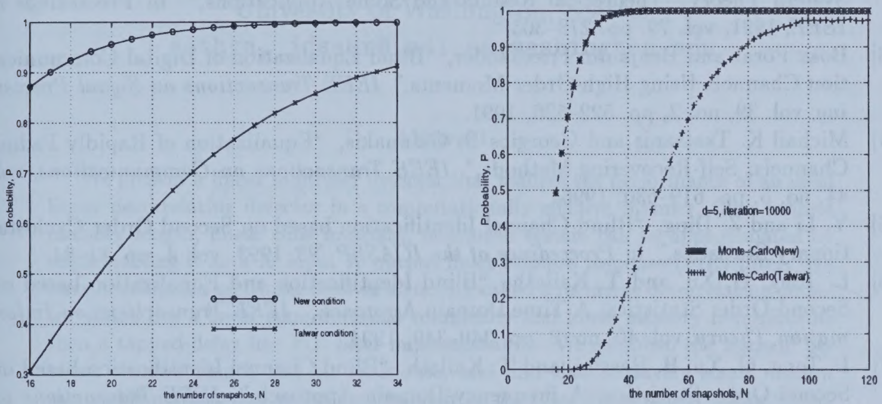
C.3 Comparison

Considering all possible combinations for the case of $d = 4$, the number of the combinations which satisfy the new identifiability condition turns out to be

$$M_{total} = M_2(N, 8) + 4M_1(N, 5) + 6M_1(N, 6) + 4M_1(N, 7) \tag{35}$$

where $M_1(N, l)$ and $M_2(N, l)$ are given by (30) and (32), respectively. This is shown in Fig.1(a).

Fig.1(b) shows an example of Monte-Carlo simulation for these probabilities when $d = 5$.



(a) (b)

Fig. 1. Comparison of the two sufficient conditions in view of the probability of receiving all the columns which satisfy the identifiability condition: (a) exact calculation (when $d = 4$), (b) Monte-Carlo Simulation (when $d = 5$, 10000 samples per each data point).

Table 1 shows the required number of snapshots N to achieve 99% probability with respect to d. Note that the number increases very fast for Talwar *et al.*'s condition whereas it increases approximately linearly for our condition.

Table 1. The number of snapshots, N, to achieve 99% probability.

d	N (Our case)	N (Talwar's case)
4	26 ($p = 0.991$)	52 ($p = 0.992$)
5	36 ($p = 0.992$)	117 ($p = 0.991$)
6	54 ($p = 0.991$)	254 ($p = 0.990$)

As shown in the results, our identifiability condition is easily achieved with high probability in a few-fold samples of d whereas Talwar's condition is achieved with relatively low probability even in tenfold samples of d. Note that M is independent of the probability.

From the result it can be concluded that blind signal separation is possible with significantly fewer snapshots than previously thought.

IV. CONCLUSION

The ambiguity problem which arises in the ML-estimate of the signals in a blind BPSK signal separation problem has been investigated. Two kinds of ambiguities were discussed in a mathematical view. A new sufficient condition for the identifiability of the data matrix was introduced and we verified that the condition can be satisfied with a reasonable number of snapshots. The simulation results confirmed that the new identifiability condition requires far fewer snapshots than the one proposed in [12].

REFERENCES

[1] G. Giannakis and J. Mendel, "Identification of Non-minimum Phase Systems using Higher-Order Statistics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 360–377, 1989.

[2] J. Mendel, "Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications," in *Proceedings of IEEE*, 1991, vol. 79, pp. 278–305.

[3] Boaz Porat and Benjamin Friedlander, "Blind Equalization of Digital Communication Channels Using High-Order Moments," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 522–526, 1991.

[4] Michail K. Tsatsanis and Georgios B. Giannakis, "Equalization of Rapidly Fading Channels: Self-Recovering Methods," *IEEE Transactions on Communications*, vol. 44, no. 5, pp. 619–630, 1996.

[5] Y. Li and Z. Ding, "Blind Channel Identification based on Second Order Cyclostationary Statistics," in *Proceedings of the ICASSP '93*, 1993, vol. 4, pp. 81–84.

[6] L. Tong, G. Xu, and T. Kailath, "Blind Identification and Equalization based on Second-Order Statistics: A Time Domain Approach," *IEEE Transactions on Imformation Theory*, vol. 40, no. 2, pp. 340–349, 1994.

[7] L. Tong, G. Xu, B. Hassibi, and T. Kailath, "Blind Channel Identification based on Second-Order Statistics: A Frequency-Domain Approach," *IEEE Transactions on Imformation Theory*, vol. 41, no. 1, pp. 329–333, 1995.

[8] Eric Moulines, Pierre Duhamel, Jean-Francois Cardoso, and Sylvie Mayrargue, "Subspace Methods for the Blind Identification of Multichannel FIR Filters," *IEEE Transactions on Signal Processing*, vol. 43, no. 2, pp. 516–525, 1995.

[9] Yingbo Hua, "Fast Maximum Likelihood for Blind Identification of Multiple FIR Channels," *IEEE Transactions on Signal Processing*, vol. 44, no. 3, pp. 661–672, 1996.

[10] A. Swindlehurst and J. Yang, "Using Least Squares to Improve Blind Signal Copy Performance," *IEEE Signal Processing Letters*, vol. 1, no. 5, pp. 80–82, 1994.

[11] Hui Liu and Guanghan Xu, "A Deterministic Approach to Blind Symbol Estimation," *IEEE Signal Processing Letters*, vol. 1, no. 12, pp. 205–207, 1994.

[12] Shilpa Talwar, Mats Viberg, and A. Paulraj, "Blind Separation of Synchronous Co-Channel Digital Signals Using an Antenna Array — Part I: Algorithms," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1184–1197, 1996.

[13] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, pp. 4–24, 1988.

[14] P. Stoica and K. C. Sharman, "Maximum Likelihood Methods for Direction-of-Arrival Estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1132–, 1990.

A Recursive Decorrelating Detector for an Asynchronous Multiuser Communication System

Sathyadev V. Uppala and John D. Sahr
Department of Electrical Engineering
University of Washington, Seattle
`sathya, jdsahr@isdl.ee.washington.edu`

Abstract

We present a linear multiuser detector that realizes the performance of an ideal linear decorrelating detector in a computationally efficient manner. This infinite memory length, linear time invariant, noncausal system can be approximated by implementing it as a K input K output finite impulse response (FIR) filter. We have developed a multiuser receiver structure called a recursive decorrelating detector which has a lower computational complexity and lower memory requirements then a tapped-delay line FIR filter implementation. We establish the tradeoff between window length and bit error rate, and find that relatively small window size yields performance indistinguishable from the ideal linear decorrelating detector. This new receiver structure can also be used to implement a minimum mean squared error (MMSE) multiuser receiver.

1 Introduction

When several users asynchronously transmit finite length messages, the linear decorrelating detector (LDD) [1] eliminates the crosstalk among the users. This receiver cannot be implemented in practice because all the users' messages must be decoded en masse, and the decoding cannot be started until all the messages are complete. Xie *et al* [2] have exploited the structure inherent in the implied equations, and have reduced the computational burden. Nevertheless, the time delay and computational burden of the LDD are still very large, motivating the development of several suboptimal receiver architectures [2, 3, 4, 5].

For the case of infinite length sequences, Lupas and Verdu [1] have shown that the LDD may be implemented as K input K output, non causal, linear time invariant, infinite memory length system. For a stable system the impulse response $h(n) \to 0$ as $|n| \to \infty$. Truncating the impulse response leads to an finite impulse response (FIR) filter implementation. This idea was used by Juntti and Aazhang [6] and they proposed the truncated LDD with tapped-delay line finite memory implementation.

In this paper we introduce an efficient receiver structure which also truncates of the impulse response for infinite length messages. The algorithm relies upon an iterative method for solving block tridiagonal system of equations known as the "incomplete cyclic odd-even reduction technique" [7]. We call the result the recursive decorrelating detector (RDD). The RDD has two principal strengths: it is computationally efficient compared to an equivalent tapped-delay line FIR filter implementation; it also provides a convenient theoretical tool for determining the appropriate window length.

In Section 2 we describe the multiuser system. In Section 3 we outline the cyclic reduction technique. In Section 4 we present the recursive decorrelating detector and in Section 5 we analyze its performance and provide design criteria for selecting the order of the RDD. We conclude with some simulation results.

2 System Model Description

We consider a multiuser system with K users. User k transmits binary data sequences $b_k(i) = \pm 1$ of length N which modulate a signature waveform $s_k(t)$ unique to that user; each bit of the messages lasts T seconds. The receiver detects the users asynchronously, each having a delay τ_k, and an amplitude $a_k(i)$ which could vary depending upon the channel; the received signal $r(t)$ is as follows.

$$r(t) = \sum_{i=1}^{N} \sum_{k=1}^{K} a_k(i) b_k(i) s_k(t - iT - \tau_k) + n(t) \tag{1}$$

The signature waveforms $s_k(t)$ have unit energy and are zero outside the interval $[0, T]$. The received signal is corrupted by zero mean white Gaussian noise $n(t)$ of variance σ.

The received signal is processed by a bank of matched filters (impulse response $s_k(-t)$) and then sampled at bit times $iT + \tau_k$. The output of the matched filter bank in vector notation is

$$\boldsymbol{y}_i = \boldsymbol{R}_{+1} \boldsymbol{A}_{i-1} \boldsymbol{b}_{i-1} + \boldsymbol{R}_0 \boldsymbol{A}_i \boldsymbol{b}_i + \boldsymbol{R}_{-1} \boldsymbol{A}_{i+1} \boldsymbol{b}_{i+1} + \boldsymbol{n}_i \tag{2}$$

The vector $\boldsymbol{b}_i = [b_1(i), \ldots, b_K(i)]^T$, $\boldsymbol{y}_i = [y_1(i), \ldots, y_K(i)]^T$ and $\boldsymbol{n}_i = [n_1(i), \ldots, n_K(i)]^T$, refer to the bits, matched filter outputs and the noise in the matched filter outputs in the ith bit interval. The diagonal matrix $\boldsymbol{A}_i = \mathrm{diag}[a_1(i), \ldots, a_K(i)]^T$ contains the amplitudes in the ith interval. The elements $r_{lk}(m)$ of the correlation matrix $\boldsymbol{R}_m$ and the noise $z_k(i)$ are given by

$$r_{lk}(m) = \int_{-\infty}^{\infty} s_l(t - \tau_l) s_k(t + mT - \tau_k)\, dt \tag{3}$$

$$n_k(i) = \int_{-\infty}^{\infty} n(t) s_k(t - iT - \tau_k)\, dt \tag{4}$$

The $\boldsymbol{n}_i$ is zero mean Gaussian with correlation given by, $E\{\boldsymbol{n}_i \boldsymbol{n}_j^T\} = \sigma^2 \boldsymbol{R}_{i-j}$.

We now construct the supervectors, $\boldsymbol{b} = [\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_N^T]^T$, of all the bits, $\boldsymbol{y} = [\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T]^T$, of all the matched filter outputs and $\boldsymbol{n} = [\boldsymbol{n}_1^T, \ldots, \boldsymbol{n}_N^T]^T$, of all the noise. We also construct the correlation matrix

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_0 & \boldsymbol{R}_{-1} & 0 & \cdots & 0 \\ \boldsymbol{R}_{+1} & \boldsymbol{R}_0 & \boldsymbol{R}_{-1} & & \vdots \\ 0 & \boldsymbol{R}_{+1} & \boldsymbol{R}_0 & \ddots & 0 \\ \vdots & & \ddots & \ddots & \boldsymbol{R}_{-1} \\ 0 & \cdots & 0 & \boldsymbol{R}_{+1} & \boldsymbol{R}_0 \end{bmatrix} \stackrel{\text{def}}{=} \mathrm{trid}_N(\boldsymbol{R}_{+1}, \boldsymbol{R}_0, \boldsymbol{R}_{-1}) \tag{5}$$

and the diagonal matrices $\boldsymbol{A} = \mathrm{diag}[\boldsymbol{A}_1, \dots, \boldsymbol{A}_N] \overset{\mathrm{def}}{=} \mathrm{diag}_N(\boldsymbol{A}_i)$ of amplitudes. The following vector equation can then be used to represent the output of the matched filter over all bit intervals.

$$\boldsymbol{y} = \boldsymbol{R}\boldsymbol{A}\boldsymbol{b} + \boldsymbol{n} \tag{6}$$

The matrix $\boldsymbol{R}$ is block tridiagonal, block Toeplitz and symmetric positive definite [1]. The linear decorrelating detector (LDD) removes the MAI by transforming the vector $\boldsymbol{y}$ to $\boldsymbol{x} = \boldsymbol{R}^{-1}\boldsymbol{y}$. The bit decision $\hat{\boldsymbol{b}}$ is obtained by thresholding $\boldsymbol{x}$ i.e, $\hat{\boldsymbol{b}} = \mathrm{sgn}\,\boldsymbol{x}$. The vector $\boldsymbol{x}$ results from solving $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$.

3 Cyclic Odd-Even Reduction Algorithm

The cyclic reduction algorithm [7] rapidly solves the linear system of equations represented by $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$, where $\boldsymbol{R}$ has a block tridiagonal structure, with invertible diagonal blocks. Matrix $\boldsymbol{R}$ is partitioned into N^2 square submatrices each of size K.

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{F}_1 & 0 & \dots & 0 \\ \boldsymbol{E}_2 & \boldsymbol{D}_2 & \boldsymbol{F}_2 & & \vdots \\ 0 & \boldsymbol{E}_3 & \boldsymbol{D}_3 & \ddots & 0 \\ \vdots & & \ddots & \ddots & \boldsymbol{F}_{N-1} \\ 0 & \dots & 0 & \boldsymbol{E}_N & \boldsymbol{D}_N \end{bmatrix} \overset{\mathrm{def}}{=} \mathrm{trid}_N(\boldsymbol{E}_i, \boldsymbol{D}_i, \boldsymbol{F}_i) \tag{7}$$

The vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are partitioned correspondingly, and we have $\boldsymbol{x} = [\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_N^T]^T$ and $\boldsymbol{y} = [\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_N^T]^T$, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are subvectors of length K. We motivate the cyclic reduction algorithm by considering the following example. We consider the three consecutive vector equations, $(2i-1)$, $(2i)$ and $(2i+1)$ from the matrix equation $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$.

$$\begin{array}{llllllll} \text{Eq. } (2i-1): & \boldsymbol{E}_{2i-1}\boldsymbol{x}_{2i-2} & + & \boldsymbol{D}_{2i-1}\boldsymbol{x}_{2i-1} & + & \boldsymbol{F}_{2i-1}\boldsymbol{x}_{2i} & = & \boldsymbol{y}_{2i-1} \\ \text{Eq. } 2i: & \boldsymbol{E}_{2i}\boldsymbol{x}_{2i-1} & + & \boldsymbol{D}_{2i}\boldsymbol{x}_{2i} & + & \boldsymbol{F}_{2i}\boldsymbol{x}_{2i+1} & = & \boldsymbol{y}_{2i} \\ \text{Eq. } (2i+1): & \boldsymbol{E}_{2i+1}\boldsymbol{x}_{2i} & + & \boldsymbol{D}_{2i+1}\boldsymbol{x}_{2i+1} & + & \boldsymbol{F}_{2i+1}\boldsymbol{x}_{2i+2} & = & \boldsymbol{y}_{2i+1} \end{array} \tag{8}$$

The variables $\boldsymbol{x}_{2i-1}$ and $\boldsymbol{x}_{2i+1}$ are eliminated from Equation $(2i)$ to obtain the following:

$$\begin{gathered} (-\boldsymbol{E}_{2i}\boldsymbol{D}_{2i-1}^{-1}\boldsymbol{E}_{2i-1})\boldsymbol{x}_{2i-2} + (\boldsymbol{D}_{2i} - \boldsymbol{E}_{2i}\boldsymbol{D}_{2i-1}^{-1}\boldsymbol{F}_{2i-1} - \boldsymbol{F}_{2i}\boldsymbol{D}_{2i+1}^{-1}\boldsymbol{E}_{2i+1})\boldsymbol{x}_{2i} \\ +(-\boldsymbol{F}_{2i}\boldsymbol{D}_{2i+1}^{-1}\boldsymbol{F}_{2i+1})\boldsymbol{x}_{2i+2} = \boldsymbol{y}_{2i} - \boldsymbol{E}_{2i}\boldsymbol{D}_{2i-1}^{-1}\boldsymbol{y}_{2i-1} - \boldsymbol{F}_{2i}\boldsymbol{D}_{2i+1}^{-1}\boldsymbol{y}_{2i+1} \end{gathered} \tag{9}$$

The new equation is in terms of only the even indexed unknowns. Using the same procedure all the odd indexed unknowns can be eliminated from the even indexed equations, resulting in a new set of equations half as large as before. This method can than be applied recursively. If N is of the form $N = 2^m - 1$ for positive integer m, we are finally left with a small linear system of order K which is easily solved. For other values of N not of the form $2^m - 1$, we'll be left with a larger system — but in this report we will assume that $N = 2^m - 1$ with the idea that N will ultimately be quite large. The *forward elimination* sketched above can be inverted through a *backward substitution* process. As it happens, for the RDD we only need the forward elimination process which is described in its general form below.

At the pth iteration we have a set of linear equations $\boldsymbol{R}^{(p)}\boldsymbol{x}^{(p)} = \boldsymbol{y}^{(p)}$, which, upon the next iteration generate a new set of equations represented by $\boldsymbol{R}^{(p+1)}\boldsymbol{x}^{(p+1)} = \boldsymbol{y}^{(p+1)}$.

The matrix $\boldsymbol{R}^{(p)}$ is given by $\boldsymbol{R}^{(p)} = \mathrm{trid}_{N_p}(\boldsymbol{E}_i^{(p)}, \boldsymbol{D}_i^{(p)}, \boldsymbol{F}_i^{(p)})$, where $N_p = 2^{m-p} - 1$. The vectors $\boldsymbol{x}^{(p)}$ and $\boldsymbol{y}^{(p)}$ are given by $\boldsymbol{x}^{(p)} = [\boldsymbol{x}_1^{(p)}, \dots, \boldsymbol{x}_{N_p}^{(p)}]$ and $\boldsymbol{y}^{(p)} = [\boldsymbol{y}_1^{(p)}, \dots, \boldsymbol{y}_{N_p}^{(p)}]$. Equations (10)–(14) indicate the mapping between $\boldsymbol{R}^{(p)}$, $\boldsymbol{y}^{(p)}$, $\boldsymbol{x}^{(p)}$ and $\boldsymbol{R}^{(p+1)}$, $\boldsymbol{y}^{(p+1)}$, $\boldsymbol{x}^{(p+1)}$.

$$\boldsymbol{E}_i^{(p+1)} = -\boldsymbol{E}_{2i}^{(p)}(\boldsymbol{D}_{2i-1}^{(p)})^{-1}\boldsymbol{E}_{2i-1}^{(p)} \tag{10}$$

$$\boldsymbol{D}_i^{(p+1)} = \boldsymbol{D}_{2i}^{(p)} - \boldsymbol{E}_{2i}^{(p)}(\boldsymbol{D}_{2i-1}^{(p)})^{-1}\boldsymbol{F}_{2i-1}^{(p)} - \boldsymbol{F}_{2i}^{(p)}(\boldsymbol{D}_{2i+1}^{(p)})^{-1}\boldsymbol{E}_{2i+1}^{(p)} \tag{11}$$

$$\boldsymbol{F}_i^{(p+1)} = -\boldsymbol{F}_{2i}^{(p)}(\boldsymbol{D}_{2i+1}^{(p)})^{-1}\boldsymbol{F}_{2i+1}^{(p)} \tag{12}$$

$$\boldsymbol{x}_i^{(p+1)} = \boldsymbol{x}_{2i}^{(p)} \tag{13}$$

$$\boldsymbol{y}_i^{(p+1)} = \boldsymbol{y}_{2i}^{(p)} - \boldsymbol{E}_{2i}^{(p)}(\boldsymbol{D}_{2i-1}^{(p)})^{-1}\boldsymbol{y}_{2i-1}^{(p)} - \boldsymbol{F}_{2i}^{(p)}(\boldsymbol{D}_{2i+1}^{(p)})^{-1}\boldsymbol{y}_{2i+1}^{(p)} \tag{14}$$

The forward elimination begins by setting $\boldsymbol{R}^{(0)} = \boldsymbol{R}$, $\boldsymbol{x}^{(0)} = \boldsymbol{x}$ and $\boldsymbol{y}^{(0)} = \boldsymbol{y}$. Since $N = 2^m - 1$ the total number of iterations are m and $p = 0, \dots, (m-1)$. At the pth iteration the vector $\boldsymbol{x}^{(p)}$ contains those subvectors of $\boldsymbol{x}$ whose indices are multiples of 2^p. Once the forward elimination is completed, an equation of the form $\boldsymbol{R}^{(m-1)}\boldsymbol{x}^{(m-1)} = \boldsymbol{y}^{(m-1)}$ remains, where $\boldsymbol{R}^{(m-1)} = \boldsymbol{D}_1^{(m-1)}$ is K by K, $\boldsymbol{x}_1^{(m-1)} = \boldsymbol{x}_{(N+1)/2}$, and $\boldsymbol{y}_1^{(m-1)} = \boldsymbol{y}_{(N+1)/2}$. The back substitution is described by Heller, but we only need the vector $\boldsymbol{x}_{(N+1)/2}$.

3.1 Incomplete Cyclic Reduction Algorithm

Heller [7] proposed the incomplete cyclic reduction algorithm to obtain an approximate solution $\hat{\boldsymbol{x}}$ for $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$, motivated by the idea that a tridiagonal matrix describes interaction only of nearest neighbors, with the influence of distant neighbors being quite small. Thus one simply stops the cyclic reduction algorithm at some iteration q and solves the equation $\boldsymbol{R}^{(q)}\boldsymbol{x}^{(q)} = \boldsymbol{y}^{(q)}$, by approximating $\boldsymbol{R}^{(q)}$ by $\mathrm{diag}_{N_q}(\boldsymbol{D}_i^{(q)})$. Thus we have $\boldsymbol{x}^{(q)} \approx \hat{\boldsymbol{x}}^{(q)} = \mathrm{diag}_{N_q}[(\boldsymbol{D}_i^{(q)})^{-1}]\boldsymbol{y}^{(q)}$. The vector $\hat{\boldsymbol{x}}^{(q)}$ provides a partial solution to the original set of equations and the complete solution $\hat{\boldsymbol{x}}$ is obtained by a backward substitution process. Heller analyzed the stability of the cyclic reduction and the conditions under which the incomplete reduction provides a good approximation to the exact solution.

We introduce the tridiagonal matrix $\boldsymbol{B}^{(p)} = \mathrm{trid}_{N_p}[(\boldsymbol{D}_i^{(p)})^{-1}\boldsymbol{E}_i^{(p)}, 0, (\boldsymbol{D}_i^{(p)})^{-1}\boldsymbol{F}_i^{(p)}]$, with the matrix infinity norm as $\|\boldsymbol{M}\|_\infty = \max_{i=1,\dots,m}\sum_{j=1}^{n}|m_{ij}|$, where $\boldsymbol{M}$ is a matrix of size m by n. Heller's conclusions about stability [7] are restated here.

Theorem 1. *If $\boldsymbol{R}^{(0)}$ is a nonsingular matrix with invertible diagonal blocks $\boldsymbol{D}_i^{(0)}$ and $\|\boldsymbol{B}^{(0)}\|_\infty < 1$ then*

1. *Cyclic Reduction is well defined, i.e. all $(\boldsymbol{D}_i^{(p)})^{-1}$ exist for $p = 0, \dots, (m-1)$ and $i = 1, \dots, N_p$.*

2. $\|\boldsymbol{B}^{(p+1)}\|_\infty \le \|(\boldsymbol{B}^{(p)})^2\|$, $p = 0, \dots, m-2$

The theorem states that the iterated block tridiagonal system rapidly approaches a simple block diagonal system if the off diagonal elements become suitably small.

4 Recursive Decorrelating Detector

Now we apply incomplete cyclic reduction algorithm to the linear decorrelating detector to develop the recursive decorrelator detector. The matrix $\boldsymbol{R}$ is defined by the correlations of the signature waveforms and $\boldsymbol{y}$ by the matched filter outputs. The output $\boldsymbol{x}$ of the LDD can be obtained by using the cyclic reduction algorithm. Since the correlation matrix (Equation (5)), is block Toeplitz, we drop the subscript i and represent $\boldsymbol{R}$ as $\mathrm{trid}_N(\boldsymbol{E}, \boldsymbol{D}, \boldsymbol{F})$, where, $\boldsymbol{E} = \boldsymbol{R}_{+1}$, $\boldsymbol{D} = \boldsymbol{R}_0$, $\boldsymbol{F} = \boldsymbol{R}_{-1}$ and $\boldsymbol{F} = \boldsymbol{E}^T$. The following proposition is true for a symmetric block Toeplitz matrix.

Proposition 1. *If $\boldsymbol{R}^{(0)}$ is symmetric block Toeplitz the cyclically reduced matrix $\boldsymbol{R}^{(p)}$ is also symmetric block Toeplitz.*

Proof. The forward elimination process (Equations (10)–(14)) preserves Toeplitz structure. □

At the pth iteration we have $\boldsymbol{R}^{(p)} = \mathrm{trid}_{N_p}(\boldsymbol{E}^{(p)}, \boldsymbol{D}^{(p)}, \boldsymbol{F}^{(p)})$ with $\boldsymbol{F}^{(p)} = (\boldsymbol{E}^{(p)})^T$. The forward elimination process for the symmetric block Toeplitz system is as follows.

$$\boldsymbol{E}^{(p+1)} = -\boldsymbol{E}^{(p)}(\boldsymbol{D}^{(p)})^{-1}\boldsymbol{E}^{(p)} \tag{15}$$

$$\boldsymbol{D}^{(p+1)} = \boldsymbol{D}^{(p)} - \boldsymbol{E}^{(p)}(\boldsymbol{D}^{(p)})^{-1}\boldsymbol{F}^{(p)} - \boldsymbol{F}^{(p)}(\boldsymbol{D}^{(p)})^{-1}\boldsymbol{E}^{(p)} \tag{16}$$

$$\boldsymbol{x}_i^{(p+1)} = \boldsymbol{x}_{2i}^{(p)} \tag{17}$$

$$\boldsymbol{y}_i^{(p+1)} = \boldsymbol{y}_{2i}^{(p)} - \boldsymbol{E}^{(p)}(\boldsymbol{D}^{(p)})^{-1}\boldsymbol{y}_{2i-1}^{(p)} - \boldsymbol{F}^{(p)}(\boldsymbol{D}^{(p)})^{-1}\boldsymbol{y}_{2i+1}^{(p)} \tag{18}$$

Similarly, we have $\boldsymbol{B}^{(p)} = \mathrm{trid}_{N_p}[(\boldsymbol{D}^{(p)})^{-1}\boldsymbol{E}^{(p)},\ \boldsymbol{0},\ (\boldsymbol{D}^{(p)})^{-1}\boldsymbol{F}^{(p)}]$. The correlation matrix $\boldsymbol{R}$ is symmetric positive definite (spd) but $\|\boldsymbol{B}^{(0)}\|_\infty$ is not necessarily less then one. The cyclic reduction algorithm is stable for any spd matrix [8]. In realistic multiuser systems we encounter $\boldsymbol{R}$ matrices whose $\|\boldsymbol{B}^{(p)}\|_\infty$ decreases with p and the quadratic convergence (Theorem 1) is achieved when the $\|\boldsymbol{B}^{(p)}\|_\infty$ dips below unity.

Using the incomplete cyclic reduction algorithm we obtain the approximate solution $\hat{\boldsymbol{x}}^{(q)} = [\mathrm{diag}_{N_q}(\boldsymbol{D}^{(q)})]^{-1}\boldsymbol{y}^{(q)}$. Because of the block diagonal matrix the solutions are decoupled i.e $\hat{\boldsymbol{x}}_i^{(q)} = (\boldsymbol{D}^{(q)})^{-1}\boldsymbol{y}_i^{(q)}$. We also note that $\hat{\boldsymbol{x}}_i^{(q)}$ maps to $\hat{\boldsymbol{x}}_{i2^q}^{(0)}$ and the bit decision $\hat{\boldsymbol{b}}_{i2^q} = \mathrm{sgn}\,\hat{\boldsymbol{x}}_{i2^q}^{(0)}$. It can be seen that $\boldsymbol{y}_i^{(q)}$ and hence $\hat{\boldsymbol{x}}_i^{(q)}$ depend only on $2^{q+1} - 1$ matched filter outputs centered around $\boldsymbol{y}_{i2^q}^{(0)}$.

The jth bit decision $\hat{\boldsymbol{b}}_j$ is obtained by processing the matched filter vector outputs in a window of length $W = 2^{q+1} - 1$, centered around $\boldsymbol{y}_j$, i.e $[\boldsymbol{y}_{j-2^q+1}, \cdots, \boldsymbol{y}_j, \cdots, \boldsymbol{y}_{j+2^q-1}]$. The processing is done by the forward elimination procedure in Equations (17) and (18). The coefficients $\boldsymbol{E}^{(p)}$ and $\boldsymbol{D}^{(p)}$ are precomputed using Equations (15) and (16). At the end of cyclic reduction the jth bit decision is given by $\hat{\boldsymbol{b}}_j^{(0)} = \mathrm{sgn}\,\hat{\boldsymbol{x}}_j^{(0)}$. A similar process is followed for obtaining the next bit decision $\hat{\boldsymbol{b}}_{j+1}$. We call this detector as the qth order recursive decorrelating detector (RDD). The computation tree in Figure 1 illustrates the process of forward elimination for a 2nd order RDD. The empty circles represent vectors with odd indices and the filled circles represent vectors with even indices. At each step (p) the interference due to the odd indexed variables is eliminated from the even indexed variables. The dashed line is the computation tree for obtaining the next bit decision. The computation tree need not be reconstructed for each bit decision because the trees for odd or even indexed decisions have lot of common nodes.

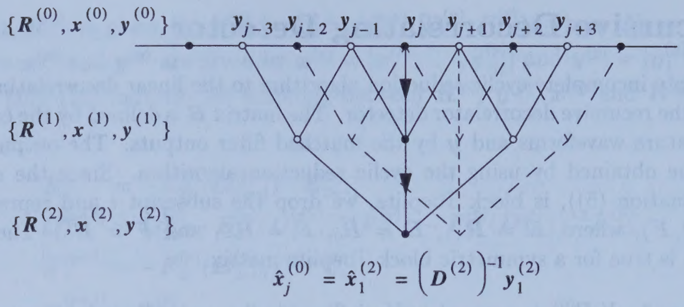

$\{R^{(0)}, x^{(0)}, y^{(0)}\}$
$\{R^{(1)}, x^{(1)}, y^{(1)}\}$
$\{R^{(2)}, x^{(2)}, y^{(2)}\}$
y_{j-3} y_{j-2} y_{j-1} y_j y_{j+1} y_{j+2} y_{j+3}
$\hat{x}_j^{(0)} = \hat{x}_1^{(2)} = \left(D^{(2)}\right)^{-1} y_1^{(2)}$

Figure 1: Computation tree for the process of forward elimination used by the qth order ($q = 2$) recursive decorrelating detector (RDD).

5 Performance Analysis and Design of the RDD

In this section we obtain an expression for the probability of error of the qth order recursive decorrelating detector. We then bound the error rate and show how the probability of error changes with q and can be made arbitrary close to that of the linear decorrelating detector. This enables us to choose the order of the RDD based on the system parameters. In order to facilitate the computation of the probability of error we prove the following proposition.

Proposition 2. *If $\boldsymbol{y}^{(0)} = \boldsymbol{R}^{(0)}\boldsymbol{A}^{(0)}\boldsymbol{b}^{(0)} + \boldsymbol{n}^{(0)}$ and $\boldsymbol{n}^{(0)}$ is Gaussian with zero mean and correlation $\sigma^2\boldsymbol{R}^{(0)}$, then the cyclically reduced vectors $\boldsymbol{y}^{(p)}$ can be represented as $\boldsymbol{y}^{(p)} = \boldsymbol{R}^{(p)}\boldsymbol{A}^{(p)}\boldsymbol{b}^{(p)} + \boldsymbol{n}^{(p)}$, where $\boldsymbol{n}^{(p)}$ is Gaussian with zero mean and correlation $\sigma^2\boldsymbol{R}^{(p)}$. The vector $\boldsymbol{b}^{(p)}$ is constructed from the subvectors of $\boldsymbol{b}^{(0)}$ whose indices are multiples of 2^p and the matrix $\boldsymbol{A}^{(p)}$ is constructed from the diagonal blocks of $\boldsymbol{A}^{(0)}$ whose indices are multiples of 2^p.*

Proof. Expressing $\boldsymbol{y}^{(p+1)}$ in terms of $\boldsymbol{y}^{(p)}$ using Equations (18), the proposition follows. ☐

Using this proposition we can immediately compute the probability of error of the qth order RDD.

Proposition 3. *The probability of error $P_{(i,k)}$ of the kth user's ith bit for a qth order recursive decorrelating detector is given by*

$$P_{(i,k)} = 2^{-2K} \sum_{\boldsymbol{b}_{i-1},\boldsymbol{b}_{i+1}\in\{-1,1\}^K} Q\left(\frac{a_k(i) - (\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{A}_{i-1}\boldsymbol{b}_{i-1})_k - (\boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{A}_{i+1}\boldsymbol{b}_{i+1})_k}{\sigma\sqrt{(\boldsymbol{D}^{-1})_{kk}}}\right) \quad (19)$$

The $\boldsymbol{D}$, $\boldsymbol{E}$, $\boldsymbol{F}$, $\boldsymbol{A}_i$ and $\boldsymbol{b}_i$ are the corresponding parameters of the qth order system.

Proof. This can be established by noting that $\boldsymbol{y}^{(q)} = \boldsymbol{R}^{(q)}\boldsymbol{A}^{(q)}\boldsymbol{b}^{(q)} + \boldsymbol{n}^{(q)}$ (from Proposition 2) and, $\hat{\boldsymbol{x}}_i = \boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{A}_{i-1}\boldsymbol{b}_{i-1} + \boldsymbol{A}_i\boldsymbol{b}_i + \boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{A}_{i+1}\boldsymbol{b}_{i+1} + \boldsymbol{D}^{-1}\boldsymbol{n}_i$. The probability of error follows from $P_{(i,k)} = \mathrm{P}(\hat{\boldsymbol{x}}_{(i,k)} < 0 \mid \boldsymbol{b}_{(i,k)} = +1)$. ☐

The probability of error $P_{(i,k)}$ depends on the relative strengths of the users' signals, the SNR of the user and noise variance. We define the dynamic range (DR) as the ratio of the powers of the strongest and weakest signals as follows.

$$\mathrm{DR} = \max_{\forall i,j,k,l} \frac{a_l(j)^2}{a_k(i)^2} \tag{20}$$

We now bound the probability of error. The bounds are expressed in terms of the dynamic range and SNR of the signals, and $\|\boldsymbol{B}^{(q)}\|_\infty$. These bounds enable us to choose the order of the RDD q based on the system requirements.

Proposition 4. *The probability of error of a qth order RDD satisfies the following bounds*

$$Q\left(\sqrt{SNR_{\mathit{eff}}}\right) \leq P_{(i,k)} \leq \frac{1}{2}\left[Q\left(\sqrt{SNR_{\mathit{eff}}}(1-\alpha)\right) + Q\left(\sqrt{SNR_{\mathit{eff}}}(1+\alpha)\right)\right] \tag{21}$$

where $\alpha = \|\boldsymbol{B}\|_\infty DR$ *and* $SNR_{\mathit{eff}} = a_k(i)^2/(\sigma^2(\boldsymbol{D}^{-1})_{kk})$ *are defined at the qth iteration.*

Proof. We cast $P_{(i,k)}$ from Equation (19) in the following form

$$P_{(i,k)} = 2^{-2K} \sum_{\boldsymbol{b}_{i-1}, \boldsymbol{b}_{i+1} \in \{-1,1\}^K} Q\left(\sqrt{\mathrm{SNR}_{\mathrm{eff}}}\,(1-\beta)\right) \tag{22}$$

$$\text{where, } \mathrm{SNR}_{\mathrm{eff}} = a_k(i)^2/(\sigma^2(\boldsymbol{D}^{-1})_{kk})$$

$$\text{and } \beta = \frac{(\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{A}_{i-1}\boldsymbol{b}_{i-1})_k + (\boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{A}_{i+1}\boldsymbol{b}_{i+1})_k}{a_k(i)}$$

The above summation consists of terms of the form $[Q(x(1-\beta)) + Q(x(1+\beta))]$ where $x = \sqrt{\mathrm{SNR}_{\mathrm{eff}}}$ and β takes on different values depending on the bit vectors $\boldsymbol{b}_{i-1}$ and $\boldsymbol{b}_{i+1}$. Since

$$Q(x) \leq \frac{1}{2}\left[Q(x(1-\alpha)) + Q(x(1+\alpha))\right] \quad \text{if } x \geq 0 \tag{23}$$

it can be seen that the probability of error $P_{(i,k)}$ is bounded below by $Q(\sqrt{\mathrm{SNR}_{\mathrm{eff}}})$.

The upper bound is expressed through the parameter β as follows.

$$\beta = \frac{1}{a_k(i)}\left(\left[\boldsymbol{D}^{-1}\boldsymbol{E} \quad \boldsymbol{D}^{-1}\boldsymbol{F}\right]\left[\boldsymbol{A}_{i-1}\boldsymbol{b}_{i-1} \quad \boldsymbol{A}_{i+1}\boldsymbol{b}_{i+1}\right]^T\right)_k \tag{24}$$

$$\Rightarrow |\beta| \leq \|\boldsymbol{B}\|_\infty \max_{\forall j,l} \frac{a_l(j)}{a_k(i)} \tag{25}$$

$$\leq \|\boldsymbol{B}\|_\infty \sqrt{\mathrm{DR}} \tag{26}$$

Defining $\alpha = \|\boldsymbol{B}\|_\infty \sqrt{\mathrm{DR}}$, we have the following inequality.

$$Q(x(1-\beta)) + Q(x(1+\beta)) \leq Q(x(1-\alpha)) + Q(x(1+\alpha)) \quad \text{if } |\alpha| \geq |\beta| \text{ and } x \geq 0 \tag{27}$$

This establishes the upper bound on $P_{(i,k)}$. ☐

In Figure 2 we have plotted the bounds on $P_e = P_{(i,k)}$. When $\alpha = 0.01$ the upper bound is indistinguishable from the lower bound. The parameter α depends on DR and $\|\boldsymbol{B}\|_\infty$, which is a function of q. The criterion for selecting the order of the RDD is the following:

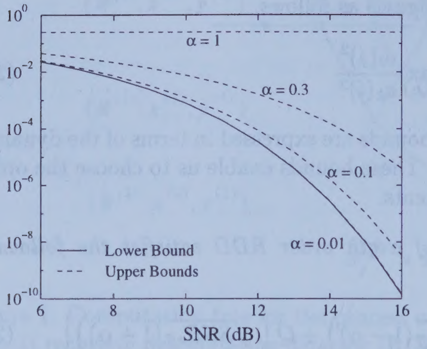


Figure 2: Bounds on the P_e of the RDD vs. $\mathrm{SNR}_{\mathrm{eff}}$. The upper bounds depend on $\alpha = ||\boldsymbol{B}||_\infty \mathrm{DR}$. When $\alpha = 0.01$, the upper bound is indistinguishable from the lower bound.

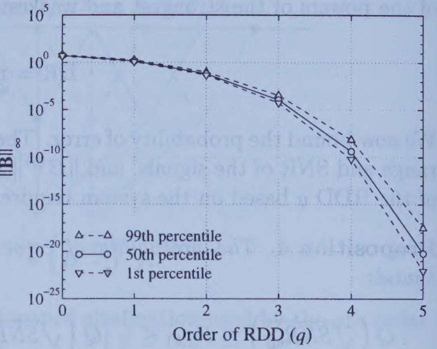


Figure 3: $||\boldsymbol{B}||_\infty$ vs. q. The number of users, $K = 63$, and number of chips, $N_c = 63$ (gold codes). $||\boldsymbol{B}||_\infty$ computed for 5×10^5 instances of the delays τ_k, uniformly distributed over $[0, T)$.

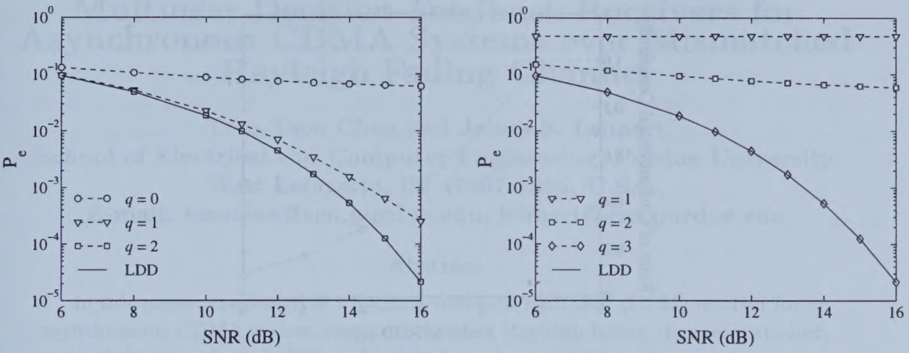
Design Criterion: Pick a q such that for a given dynamic range DR, the α is small, e.g. $\alpha = 0.01$.

We illustrate the choice of q for a multiuser system with with $K = 63$ users. The signature waveforms are gold codes with $N_c = 63$ chips per bit. We compute the $||\boldsymbol{B}||_\infty$ for this system for different values of q. The results are shown in Figure 3. We observe that $||\boldsymbol{B}||_\infty$ decreases quadratically with q, certainly when $||\boldsymbol{B}||_\infty < 1$. If $\mathrm{DR} = 0$ dB and we want an $\alpha = 0.01$ then $||\boldsymbol{B}||_\infty$ should be around 0.01. Figure 3 shows that choosing $q = 2$ ensures this. Similarly if $\mathrm{DR} = 40$ or 80 dB then q should be 3 or 4 respectively.

6 Simulation Results and Discussion

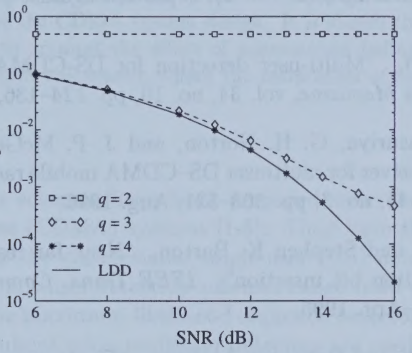
In this section we compare the performance of the RDD and ideal LDD for a system of $K = 63$ users using 63 chip gold codes. The gold codes were generated using m-sequences which had the characteristic polynomial coefficients (1000011) and (1010111). The error probabilities were averaged over 500 instances of random delays uniformly distributed over the interval $[0, T)$. For a specific set of delays the error probabilities were computed by a technique known as the saddle point integration method [9, 10, 11]. In Figures 4(a) to (c) we plot the probability of error of user number 1 versus the SNR per bit for different dynamic ranges. The SNR per bit is defined as $\mathrm{SNR} = 10\log(a_k(i)^2/\sigma^2)$. It can be seen that a higher order RDD is needed to achieve the performance of a LDD as the dynamic range increases. For $\mathrm{DR} = 0$ dB, 40 dB and 80 dB choices of $q = 2$, 3 and 4 (window lengths are 7, 15, and 31) achieve the performance of the infinite data length LDD [1]. This example verifies the design bounds developed in the previous section.

The RDD requires fewer computations and lesser memory than a corresponding tapped-delay line FIR filter. This is achieved by keeping track of the relevant parts of the computation trees which generate the bit decisions and noting that a new bit decision requires the computation of only one additional branch of the tree. In order to obtain the ith bit decision, the FIR filter needs $(2^{q+1} - 1)K = WK$ multiplies/add,



(a) Dynamic Range, DR = 0 dB.

(b) Dynamic Range, DR = 40 dB.



(c) Dynamic Range, DR = 80 dB.

Figure 4: P_e vs. SNR for user 1, for the RDD with different values of q and dynamic range. The number of users $K = 63$, number of chips $N_c = 63$ (gold codes).

whereas the qth order RDD needs $K(2q+1) \approx 2K \log_2 W$ multiplies/adds. The ratio of computational burden versus q (equivalent to window length) is shown in Figure 5. We also find that the ratio of the memory requirements of the RDD and the FIR filter is reasonably approximated by $(2q+1)/(2^{q+1}-1)$, for large user population and small q.

With this example we conclude our presentation of the Recursive Decorrelating Detector (RDD), which uses the cyclic reduction algorithm to rapidly and efficiently implement the linear decorrelating detector (LDD).

References

[1] Ruxandra Lupas and Sergio Verdu, "Near–far resistance of multiuser detectors in asynchronous channels", *IEEE Trans. Commun.*, vol. 38, no. 4, pp. 496–508, Apr. 1990.

[2] Zhenhua Xie, Robert T. Short, and Craig K. Rushforth, "A family of suboptimum detectors for coherent multiuser communications", *IEEE Trans. Commun.*, vol. 8, no. 4, pp. 683–690, May 1990.

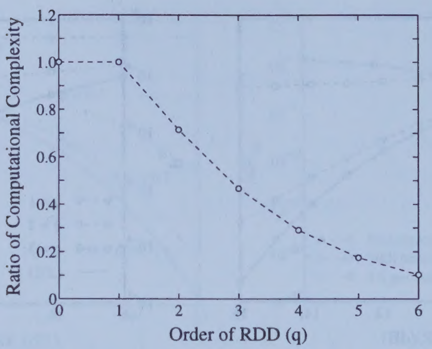


Figure 5: The ratio of computational complexity of a qth order RDD to that of a tapped-delay line FIR filter of length $(2^{q+1} - 1)$, is plotted against q.

[3] Shimon Moshavi, "Multi-user detection for DS-CDMA communications", *IEEE Communications Magazine*, vol. 34, no. 10, pp. 124–136, Oct. 1996.

[4] S. S. H. Wijayasuriya, G. H. Norton, and J. P. McGeehan, "A sliding window decorrelating receiver for multiuser DS–CDMA mobile radio networks", *IEEE Trans. Veh. Tech.*, vol. 45, no. 3, pp. 503–521, Aug. 1996.

[5] Fu Chun Zheng and Stephen K. Barton, "Near–far resistant detection of CDMA signals via isolation bit insertion", *IEEE Trans. Commun.*, vol. 43, no. 2–4, pp. 1313–1317, Feb.-Apr. 1995.

[6] Markku J. Juntti and Behnaam Aazhang, "Finite memory-length linear multiuser detector for asynchronous CDMA communications", *IEEE Trans. Commun.*, vol. 5, no. 45, pp. 611–622, May 1997.

[7] Don Heller, "Some aspects of the cyclic reduction algorithm for block tridiagonal linear systems", *SIAM J. Numer. Anal.*, vol. 13, no. 4, pp. 484–496, Sept. 1976.

[8] I. K. Abu-Shumays, "Incomplete block cyclic reduction as a preconditioner for polynomial iterative methods", *SIAM J. Sci. Stat. Comput.*, vol. 11, no. 3, pp. 545–562, May 1990.

[9] Carl W. Helstrom, "Calculating error probabilities for intersymbol and cochannel interference", *IEEE Trans. Commun.*, vol. COM. 34, no. 5, pp. 430–435, May 1986.

[10] Jack W. Stokes and James A. Ritcey, "Error probabilities of synchronous DS/CDMA systems with random and deterministic signature sequences for ideal and fading channels", in *Proc. of IEEE ICC'96*, Dallas, Tx, U.S.A., June 1996, vol. 3, pp. 1518–1522.

[11] Sathyadev V. Uppala and John D. Sahr, "Computing probability of errors of linear multiuser detectors in white gaussian noise and rayleigh fading channels", Submitted to ICC'98, Aug. 1997.

Multiuser Decision-Feedback Receivers for Asynchronous CDMA Systems over Mismatched Rayleigh Fading Channel

Tsao-Tsen Chen and James S. Lehnert
School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907-1285, U.S.A.
E-mail: tsaotsen@ecn.purdue.edu, lehnert@ecn.purdue.edu

Abstract

In this paper, the design of a decision-feedback multiuser (DFM) receiver for an asynchronous CDMA system over a mismatched Rayleigh fading channel is studied, and a tight upper bound (UB) and a tight lower bound (LB) on the pairwise symbol error probability (PSEP) are derived for some of the DFM detection schemes. The performance bounds, which are computed by efficient analytic techniques, can serve as reliable references for the CDMA system design. It is shown that the proposed system, which takes into account the effect of mismatched fading on the system design, has some gain over the system where the estimation of fading is assumed to be exact.

I Introduction

Several multiuser detection schemes have been proposed over the past few years for code-division multiple-access (CDMA) systems [1-5]. They have the advantage of performance improvement over the conventional single-user (CS) detection schemes [6-8]. The complexities of these multiuser detection schemes range from exponential with the number of users K, when the maximum-likelihood sequence detectors [1] are used, to linear with K, when various suboptimum multiuser detectors are used [2-5]. In [3], various low-complexity suboptimum multiuser receivers for convolutionally coded asynchronous CDMA systems are studied. In the decoding process, either the partitioned approach, where the equalization and decoding are separated, or the integrated approach, where the equalization and decoding are combined, is used. In particular, the nonlinear decision-feedback receivers are shown to have good performances for both the partitioned and integrated approaches, and the latter has better performance than the former at the expense of higher complexity. Hence, in this paper we focus on the nonlinear decision-feedback receiver with the partitioned approach, which has both the advantages of low-complexity and good performance. Also, we study only the part of the receiver that performs equalization, since it is central to the receiver design in the partitioned approach. Therefore, the error rate of the part that performs equalization is used as a measure of the overall system performance. We call the detection scheme studied in this paper the decision-feedback multiuser (DFM) detection scheme.

A most common form of CDMA is direct-sequence spread-spectrum multiple-access (DS/SSMA) in which each user is assigned a particular signature sequence which is modulated on the carrier along with the data [1-8]. In [8], the coded DS/SSMA system with CS detection over the mismatched Rayleigh fading channel is studied, and an arbitrarily tight UB and LB on the pairwise error probability are obtained by the proposed modified bounding technique for various decoding schemes. In practice the channel parameters are not perfectly known to the receiver, and the system performance will degrade accordingly. This is especially true for the multiuser detection scheme, and it is our goal here to investigate and reduce the effects of mismatched Rayleigh fading on the system performance.

Similar works have been reported in [4], where Gaussian approximations are used for the various estimation errors and the multiple-access interference (MAI), and [5], where a 2-user asynchronous model is used and asymptotic multiuser efficiency is investigated. In this paper, we extend the results of [8] to the case where DFM detection is used. The decision rules for various DFM detection schemes over the mismatched Rayleigh fading channel are first derived and investigated. Then, since the general two-dimensional signal set is considered, the pairwise symbol error probability (PSEP) is studied. The UB and LB on the PSEP for some of these detection schemes are then obtained via the residue method [10] and the modified bounding technique [8].

II System Model

The system model that we examine is shown in Fig. 1. There are K users sharing a common channel. We define the kth user's baseband coded signal by $b^{(k)}(t) = \sum_{j=-\infty}^{\infty} b_j^{(k)} p_T(t-jT)$, where $\{b_j^{(k)}\}$ is the output symbol sequence from the rate n/m encoder of the kth transmitter, T is the symbol interval, and $p_T(t)$ is the unit pulse on $[0,T)$. We consider the general two-dimensional signal set in this paper, and $b_j^{(k)}$ is then a complex number. The signal constellation is normalized so that the average signal energy equals 1. All signature sequences are deterministic. The signature sequence $\{\alpha_h^{(k)}\}$, where each $\alpha_h^{(k)} \in \{-1,1\}$, is used to form the kth user's spreading waveform that is given by $\alpha^{(k)}(t) = \sum_{h=-\infty}^{\infty} \alpha_h^{(k)} \psi(t-hT_c)$, where $\psi(t)$ is the common chip waveform for all signals. The chip waveform $\psi(t)$ is time limited to $[0,T_c)$, where the chip duration T_c satisfies $T_c = T/N$, and N is the period of the signature sequence. The chip waveform is also normalized such that $\int_0^{T_c} \psi^2(t)dt = T_c$. The baseband transmitted signal of the kth user is given by $S^{(k)}(t) = \sqrt{2P}\alpha^{(k)}(t)b^{(k)}(t)\exp(j\phi_k)$, where P is the power of the transmitted signal, and ϕ_k is the phase of the kth transmitter.

The received baseband signal is given by

$$r(t) = \sum_{k=1}^{K} \sqrt{2P} g^{(k)}(t)\alpha^{(k)}(t-\tau_k)b^{(k)}(t-\tau_k)\exp(j\theta_k) + n(t), \tag{1}$$

where $\{g^{(k)}(t)\}$ is a sequence of independent, zero mean, complex Gaussian (CG) fading processes with a common variance of $\sigma_g^2 = \frac{1}{2}E\{|g^{(k)}(t)|^2\}$, τ_k is the time delay parameter that accounts for propagation delay and the lack of synchronism between the transmitters, $\theta_k = \phi_k - \omega_c\tau_k$ is the phase angle of the kth carrier, ω_c is the carrier frequency, and $n(t)$ is the complex envelope of the channel white Gaussian noise with double-sided power spectral density N_o. We assume that $g^{(k)}(t) = g_j^{(k)}$ over the interval $[jT+\tau_k, (j+1)T+\tau_k)$ for $k = 1\ldots K$. We also assume that $g_j^{(k)}$'s are independently and identically distributed (iid) CG random variables (rv's) for all j and $1 \le k \le K$, each with zero mean and variance σ_g^2. The average received symbol energy and bit energy are given by $E_s = 2\sigma_g^2 PT$, and $E_b = \frac{1}{n}E_s$ respectively. The centralized receiver is considered for the DFM detection scheme. That is, the receiver consists of a bank of matched filters, and jointly demodulates the signals of all users. Also, coherent detection is considered in the receiver. We assume that the receiver has perfect knowledge of τ_k and θ_k for $k = 1\ldots K$, and of the signature sequences of all users. Without loss of generality, we assume that $\tau_1 = 0$, $0 < \tau_k < T$ for $k = 2\ldots K$, and, for notational simplicity, we consider the decision process for the 1st user in this paper. Note that for the evaluation of the average system performance, τ_k and θ_k are assumed to be uniformly distributed on $[0,T)$ and $[0,2\pi)$, respectively, for $k = 2\ldots K$. The output of the filter matched to the 1st user at time $(j+1)T + \tau_1$ is given by

$$r_j^{(1)} = u_j^{(1)} b_j^{(1)} + n_j^{(1)} + \sum_{k=2}^{K} (u_{j-1}^{(k)} I_j^{(k)} + u_j^{(k)} J_j^{(k)}) \triangleq u_j^{(1)} b_j^{(1)} + n_j^{(1)} + w_j^{(1)}, \qquad (2)$$

where $u_j^{(k)} = \sqrt{2PT} g_j^{(k)}$, $k = 1, \ldots, K$, is a zero-mean CG rv with variance $\sigma_u^2 = E_s$, and $n_j^{(1)}$ is a zero-mean CG rv with variance $\sigma_n^2 = N_0$. $I_j^{(k)}$ and $J_j^{(k)}$ are the past and future MAI [2], respectively, from the kth user, and are given by

$$I_j^{(k)} = \frac{1}{T}[b_{j-1}^{(k)} R_{k,1}(\tau_k)] \exp(j\theta_k), \qquad (3)$$

and

$$J_j^{(k)} = \frac{1}{T}[b_j^{(k)} \hat{R}_{k,1}(\tau_k)] \exp(j\theta_k), \qquad (4)$$

where $R_{k,1}(\cdot)$ and $\hat{R}_{k,1}(\cdot)$ are the continuous-time partial cross-correlation functions of the kth and 1st spreading waveforms defined in [6]. By (3) and (4), we define the MAI vector $\mathbf{I}_j$ as $\mathbf{I}_j = [I_j^{(2)}, J_j^{(2)}, \ldots, I_j^{(K)}, J_j^{(K)}]$.

For comparison, the CS detection scheme is also studied. In this case, the noncentralized receiver is used. That is, each user has only one filter that is matched to the spreading waveform of the corresponding transmitter, and demodulates the received signal independently from other users. Symmetry allows us to consider only the receiver that is listening to the first transmitter (the "desired receiver"). Coherent detection is assumed, and the 1st receiver now only has perfect knowledge of τ_1 and θ_1. Accordingly, we assume that $\tau_1 = \theta_1 = 0$, and for $k \neq 1$, the delays τ_k and the phase angles θ_k are rv's that are uniformly distributed on $[0, T)$ and $[0, 2\pi)$, respectively.

We assume that the channel estimation is available for the fading rv $u_j^{(k)}$, and is denoted as $v_j^{(k)}$ for all j and $1 \leq k \leq K$. We also assume that $v_j^{(k)}$'s are iid CG rv's with zero mean and variance of σ_v^2, and the covariance of $u_j^{(k)}$ and $v_j^{(k)}$ is σ_{uv}^2. The correlation coefficient of $u_j^{(k)}$ and $v_j^{(k)}$ is then given by $u = \sigma_{uv}^2/(\sigma_u \sigma_v)$, and is assumed to be known to the receiver. Since mismatched fading is considered, we have $u \neq 1$. The above assumptions are similar to those in [10] and [11]. However, the more general case where u is not perfectly known to the receiver is also studied in this paper. Note that for the DFM detection scheme, the centralized receiver can utilize $v_j^{(k)}$ for $k = 1 \ldots K$, and, for the CS detection scheme, the desired receiver can only utilize $v_j^{(1)}$.

III Characterization of the MAI Statistic

In this section, we characterize the MAI statistic, which is useful for the evaluation of the average system performance. We assume that τ_k's and θ_k's are independent uniform rv's on $[0, T)$ and $[0, 2\pi)$, respectively, for $k = 2 \ldots K$. Then, the modified bounding technique proposed in [8] is applied to get the key result (8). For simplicity, we study the rectangular chip waveform only. However, we can generalize the results to the arbitrary chip waveform case by applying the techniques used in [7]. We will see that the system performance of some detection schemes depends on the MAI statistic I_j defined by

$$I_j = \sum_{k=2}^{K} (|I_j^{(k)}|^2 + |J_j^{(k)}|^2) \triangleq \sum_{k=2}^{K} M_j^{(k)}, \qquad (5)$$

where $I_j^{(k)}$ and $J_j^{(k)}$ are as defined in (3) and (4). It is shown in [6] that for $0 \leq \lambda_k T_c \leq \tau_k < (\lambda_k + 1) T_c$, $R_{k,1}(\tau_k) = C_{k,1}(\lambda_k - N)(T_c - s_k) + C_{k,1}(\lambda_k - N + 1) s_k$, and $\hat{R}_{k,1}(\tau_k) = C_{k,1}(\lambda_k)(T_c - s_k) + C_{k,1}(\lambda_k + 1) s_k$, where λ_k is uniformly distributed on the set $\{0, \ldots, N-1\}$, $s_k = \tau_k - \lambda_k T_c$ is uniformly distributed on $[0, T_c)$, and $C_{k,1}(\cdot)$ is the aperiodic crosscorrelation function defined in [6]. Then $M_j^{(k)}$ can be expressed as

$$M_j^{(k)} = |A_{1,j}^{(k)} U_k + A_{2,j}^{(k)}(1-U_k)|^2 + |B_{1,j}^{(k)} U_k + B_{2,j}^{(k)}(1-U_k)|^2, \tag{6}$$

where $U_k = s_k/T_c$ is uniformly distributed on $[0,1)$, and

$$\begin{aligned} A_{1,j}^{(k)} &= \tfrac{1}{N}[b_{j-1}^{(k)} C_{k,1}(\lambda_k - N + 1)], \quad B_{1,j}^{(k)} = \tfrac{1}{N}[b_j^{(k)} C_{k,1}(\lambda_k + 1)], \\ A_{2,j}^{(k)} &= \tfrac{1}{N}[b_{j-1}^{(k)} C_{k,1}(\lambda_k - N)], \quad \text{and} \quad B_{2,j}^{(k)} = \tfrac{1}{N}[b_j^{(k)} C_{k,1}(\lambda_k)]. \end{aligned} \tag{7}$$

We define the discrete rv $D^{(k)}$, which characterizes $M_j^{(k)}$, as $D^{(k)} = i$, if $\frac{i}{N_u} \le M_j^{(k)} < \frac{i+1}{N_u}$, where N_u is a positive integer. From (6) and (7), if we condition on the set of rv's $\Lambda = \{\lambda_k, b_{j-1}^{(k)}, b_j^{(k)}; k = 2 \ldots K\}$, $M_j^{(k)}$ is a quadratic function of the rv U_k. The conditional cdf of $M_j^{(k)}$, $F_{M_j^{(k)}}(\cdot \mid \Lambda)$, is given in Appendix A of [8]. The conditional probability mass function (pmf) $d^{(k)}(\cdot \mid \Lambda)$ of $D^{(k)}$ is then given by $d^{(k)}(i \mid \Lambda) = F_{M_j^{(k)}}(\frac{i+1}{N_u} \mid \Lambda) - F_{M_j^{(k)}}(\frac{i}{N_u} \mid \Lambda)$. To obtain the (unconditional) pmf $d^{(k)}(\cdot)$ of $D^{(k)}$, we have to average $d^{(k)}(i \mid \Lambda)$ over the set Λ of rv's. In order to model the total MAI statistic, we define the discrete rv D, which characterizes I_j, as $D = \sum_{k=2}^{K} D^{(k)}$. The pmf of D is defined as $d(i) = \Pr(D = i)$. Since the rv's in the set $\{D^{(k)}; k = 2, \ldots, K\}$ are independent, $d(\cdot)$ can be computed by performing a discrete convolution of the discrete densities $d^{(k)}(\cdot)$. Let the interval $IK(i)$ be defined as

$$IK(i) = \{x; \frac{i}{N_u} \le x < \frac{i+K-1}{N_u}\}. \tag{8}$$

Then the probability mass $d(i)$ means part of the probability that I_j falls in the interval $IK(i)$. This will be used to compute the UB and LB on the PSEP in Section IV. The parameter N_u determines the length of the interval $IK(i)$, and hence controls the tightness of the bounds. By following the approach as in [8], the expected value of $M_j^{(k)}$ can be shown to be $E(M_j^{(k)}) = [2u_{k,1}(0) + u_{k,1}(1)]/(3N^3)$, where $u_{k,1}(i) = \sum_{l=1-N}^{N-1} C_{k,1}(l) C_{k,1}(l+i)$, and $C_{k,1}(\cdot)$ is the aperiodic crosscorrelation function defined in [6]. The expected value of I_j is then given by

$$E(I_j) = \sum_{k=2}^{K} E(M_j^{(k)}) \triangleq M_I. \tag{9}$$

The value M_I will be used in a suboptimum decision scheme for the CS detection in Section V.

IV Decision-Feedback Multiuser Detection

In this section, the derivation of the DFM detection schemes over the mismatched Rayleigh fading channel is given. The multistage detection algorithm (MDA) [2-5] is used in the DFM detection to enhance the system performance. In [2], the performances of several MDA's are compared, and the proposed hybrid MDA is shown to outperform other MDA's. Hence, we use the hybrid MDA in this paper. The UB and LB on the PSEP are then derived for two special cases. For the correlation coefficient u of the mismatched fading, first we assume that u is perfectly known to the receiver, and then generalize this to the case where u is not perfectly known in Section C.

A Derivation of the Decision Rule

Since DFM detection and MDA are considered in the receiver, the estimations of the past and future MAI, $I_j^{(k)}$ and $J_j^{(k)}$, at the lth stage are available and given by $\hat{I}_j^{(k)}(l)$ and $\hat{J}_j^{(k)}(l)$, respectively. Then for the 1st user at the jth interval, the channel state information (CSI) vector at the lth stage is defined as $\mathbf{V}_j(l) = [v_j^{(1)}, \hat{w}_j^{(1)}(l)]^T$, where $\hat{w}_j^{(1)}(l)$ is the estimate of $w_j^{(1)}$ (see (2)) at the lth stage and is defined as $\hat{w}_j^{(1)}(l) =$

$\sum_{k=2}^{K}[v_{j-1}^{(k)}\hat{I}_j^{(k)}(l)+v_j^{(k)}\hat{J}_j^{(k)}(l)]$. The superscript T denotes transpose. We also define the random vector $\mathbf{t}(l)$ as $\mathbf{t}(l)=[r_j^{(1)},\mathbf{V}_\mathbf{j}(l)^T]^T$. It follows that $\hat{w}_j^{(1)}(l)$ is a CG rv with zero mean and variance of $\hat{I}_j(l)\sigma_v^2$, where $\hat{I}_j(l)=\sum_{k=2}^{K}(|\hat{I}_j^{(k)}(l)|^2+|\hat{J}_j^{(k)}(l)|^2)$, and the covariance of $\hat{w}_j^{(1)}(l)$ and $w_j^{(1)}$ is given by $C(l)\sigma_{uv}^2$, where $C(l)=\sum_{k=2}^{K}\{I_j^{(k)}[\hat{I}_j^{(k)}(l)]^*+J_j^{(k)}[\hat{J}_j^{(k)}(l)]^*\}$. Since we assume that the receiver has perfect knowledge of τ_k's and θ_k's, by following the hybrid MDA [2], $\hat{I}_j^{(k)}(l)$ and $\hat{J}_j^{(k)}(l)$ can be obtained from (3) and (4) as

$$\hat{I}_j^{(k)}(l)=\frac{1}{T}[\hat{b}_{j-1}^{(k)}(l)R_{k,1}(\tau_k)]\exp(j\theta_k)\quad \text{for } l\geq 1, \tag{10}$$

and

$$\hat{J}_j^{(k)}(l)=\begin{cases} 0 & \text{if } l=1 \\ \frac{1}{T}[\hat{b}_j^{(k)}(l-1)\hat{R}_{k,1}(\tau_k)]\exp(j\theta_k) & \text{if } l>1, \end{cases} \tag{11}$$

respectively, where $\hat{b}_j^{(k)}(l)$ is the output of the decision device at the lth stage at time $(j+l-1)T+\tau_k$ and is an estimate of $b_j^{(k)}$

A.1 Optimum Decision Scheme (OD Scheme)

By conditioning on the transmitted symbol b for the 1st user at the jth interval, and assuming that the receiver has perfect knowledge of the past and future MAI, we have $\hat{I}_j(l)=C(l)=I_j$, and the CSI vector becomes $\mathbf{V}_\mathbf{j}=[v_j^{(1)},\hat{w}_j^{(1)}]^T$, where $\hat{w}_j^{(1)}=\sum_{k=2}^{K}[v_{j-1}^{(k)}I_j^{(k)}+v_j^{(k)}J_j^{(k)}]$. The pdf of $\mathbf{t}$, where $\mathbf{t}=[r_j^{(1)},\mathbf{V}_\mathbf{j}{}^T]^T$, is given by

$$p(\mathbf{t}|b)=\frac{\exp(-\frac{1}{2}\mathbf{t}^\dagger\mathbf{\Phi}^{-1}\mathbf{t})}{(2\pi)^2\det(\mathbf{\Phi})}, \tag{12}$$

where

$$\mathbf{\Phi}=\begin{bmatrix} \sigma_u^2(|b|^2+I_j)+N_o & b\sigma_{uv}^2 & I_j\sigma_{uv}^2 \\ b^*(\sigma_{uv}^2)^* & \sigma_v^2 & 0 \\ I_j(\sigma_{uv}^2)^* & 0 & I_j\sigma_v^2 \end{bmatrix} \tag{13}$$

is the correlation matrix of $\mathbf{t}$ conditioned on b. The notations det, † and $*$ denote determinant, conjugate transpose and conjugate respectively. By taking the opposite of the log of the pdf defined in (12), and substracting a common term, the decision metric of the DFM-OD scheme for the correct symbol b is given by

$$\mathrm{M}(b)=\mathbf{t}^\dagger\mathbf{\Phi}^{-1}\mathbf{t}-\frac{|\hat{w}_j^{(1)}|^2}{\sigma_v^2 I_j}+2\ln(Y_1)=\frac{|r_j^{(1)}-\beta v_j^{(1)}b-\beta\hat{w}_j^{(1)}|^2}{Y_1}+2\ln(Y_1), \tag{14}$$

where

$$\mathbf{\Phi}^{-1}=\begin{bmatrix} \frac{1}{Y_1} & \frac{-\beta b}{Y_1} & \frac{-\beta}{Y_1} \\ \frac{-\beta^* b^*}{Y_1} & |\beta|^2(\frac{|b|^2}{Y_1})+\frac{1}{\sigma_v^2} & |\beta|^2(\frac{b^*}{Y_1}) \\ \frac{-\beta^*}{Y_1} & |\beta|^2(\frac{b}{Y_1}) & \frac{|\beta|^2}{Y_1}+\frac{1}{\sigma_v^2 I_j} \end{bmatrix}, \tag{15}$$

$Y_1=\frac{\det(\mathbf{\Phi})}{\sigma_v^4 I_j}=\sigma_u^2(1-|u|^2)|b|^2+\sigma_u^2(1-|u|^2)I_j+N_o$, and $\beta=u(\sigma_u/\sigma_v)$. From (14), we see that the decision rule depends on the MAI vector $\mathbf{I}_\mathbf{j}$, which in practice is not perfectly known to the receiver. Hence, a suboptimum decision scheme is proposed in Section A.2.

A.2 Suboptimum Decision Scheme (SD Scheme)

In this section, we assume that the perfect knowledge of the MAI vector $\mathbf{I}_\mathbf{j}$ is not available to the receiver. Since DFM detection is considered in the receiver, $I_j^{(k)}$ and $J_j^{(k)}$ are approximated by the decision-feedback estimations $\hat{I}_j^{(k)}(l)$ and $\hat{J}_j^{(k)}(l)$, respectively, at the lth stage, and then I_j is in turn approximated by $\hat{I}_j(l)$. The decision metric and decision rule for the DFM-SD scheme at the lth stage are then given by

$$\mathrm{M}(b,l)=\mathbf{t}(l)^\dagger\mathbf{\Phi}^{-1}(l)\mathbf{t}(l)-\frac{|\hat{w}_j^{(1)}(l)|^2}{\sigma_v^2\hat{I}_j(l)}+2\ln(Y_1(l))=\frac{|r_j^{(1)}-\beta v_j^{(1)}b-\beta\hat{w}_j^{(1)}(l)|^2}{Y_1(l)}+2\ln(Y_1(l)), \tag{16}$$

and $\hat{b}_j^{(1)}(l) = \arg\min_{x \in \text{signal set}} \mathrm{M}(x, l)$, respectively, where

$$\mathbf{\Phi}^{-1}(l) = \begin{bmatrix} \frac{1}{Y_1(l)} & \frac{-\beta b}{Y_1(l)} & \frac{-\beta}{Y_1(l)} \\ \frac{-\beta^* b^*}{Y_1(l)} & |\beta|^2[\frac{|b|^2}{Y_1(l)}] + \frac{1}{\sigma_v^2} & |\beta|^2[\frac{b^*}{Y_1(l)}] \\ \frac{-\beta^*}{Y_1(l)} & |\beta|^2[\frac{b}{Y_1(l)}] & \frac{|\beta|^2}{Y_1(l)} + \frac{1}{\sigma_v^2 \hat{I}_j(l)} \end{bmatrix}, \tag{17}$$

and $Y_1(l) = \sigma_u^2(1 - |u|^2)|b|^2 + \sigma_u^2(1 - |u|^2)\hat{I}_j(l) + N_o$. Note that $\mathbf{\Phi}^{-1}(l)$ and $Y_1(l)$ are approximations of $\mathbf{\Phi}^{-1}$ and Y_1, respectively, at the lth stage.

B Pairwise Symbol Error Probability

We use the suboptimum decision mertic (16) in the following derivation of the PSEP. Let the transmitted symbol be b and an error symbol be $\bar{b}$ for the 1st user at the jth interval. The PSEP means the probability that the decision device at some stage chooses $\bar{b}$ instead of b, and is denoted by $\Pr(b \to \bar{b})$. In order to fairly compare the performances of different decision schemes, we define the average pairwise symbol error probability (APSEP) as

$$\Pr(b \leftrightarrow \bar{b}) = \frac{1}{2}\Pr(b \to \bar{b}) + \frac{1}{2}\Pr(\bar{b} \to b). \tag{18}$$

Note that the APSEP is the Bayes error, and the optimum decision scheme is the scheme that minimizes the APSEP, not the PSEP. However, it suffices to give the derivation of the PSEP only. The following approach is similar to those in [10] and [11]. Note that if $\bar{b}$ is conditioned instead of b, $\mathbf{\Phi}^{-1}(l)$ and $Y_1(l)$ are replaced by $\bar{\mathbf{\Phi}}^{-1}(l)$, and $\bar{Y}_1(l)$, respectively. The decision device at the lth stage will choose $\bar{b}$ instead of b if $\mathrm{M}(\bar{b}, l) < \mathrm{M}(b, l)$, or $\mathcal{D}(l) < \delta(l)$, where $\mathcal{D}(l) = \mathbf{t}(l)^\dagger \mathbf{F}(l)\mathbf{t}(l)$, $\mathbf{F}(l)$ is the difference between $\bar{\mathbf{\Phi}}^{-1}(l)$ and $\mathbf{\Phi}^{-1}(l)$, and $\delta(l) = 2\ln[Y_1(l)/\bar{Y}_1(l)]$. Note that, by conditioning on the MAI vector $\mathbf{I}_\mathrm{j}$, $\mathcal{D}(l)$ is a Hermitian quadratic form of the conditional CG random vector $\mathbf{t}(l)$, where the conditional correlation matrix of $\mathbf{t}(l)$ is given by

$$\mathbf{R}(l) = \begin{bmatrix} \sigma_u^2(|b|^2 + I_j) + N_o & b\sigma_{uv}^2 & C(l)\sigma_{uv}^2 \\ b^*(\sigma_{uv}^2)^* & \sigma_v^2 & 0 \\ C(l)^*(\sigma_{uv}^2)^* & 0 & \hat{I}_j(l)\sigma_v^2 \end{bmatrix}. \tag{19}$$

By following the approach as in [12], the two-sided Laplace transform of the conditional pdf of $\mathcal{D}(l)$ can be shown to be

$$\Phi_{\mathcal{D}(l)|\mathbf{I}_\mathrm{j}}(s) = \frac{p_1 p_2}{(s - p_1)(s - p_2)}, \tag{20}$$

where $p_1 = w - \sqrt{w^2 - p_1 p_2}$ and $p_2 = w + \sqrt{w^2 - p_1 p_2}$ are the left plane and right plane poles, respectively, with

$$p_1 p_2 = -\frac{Y_1(l)\bar{Y}_1(l)}{|u|^2\sigma_u^2|b - \bar{b}|^2 Z_1(l)}, \tag{21}$$

$$w = \frac{(1 - |u|^2)(|b|^2 - |\bar{b}|^2)Z_1(l) + |u|^2|b - \bar{b}|^2 Y_1(l)}{2|u|^2|b - \bar{b}|^2 Z_1(l)}, \tag{22}$$

and $$Z_1(l) = \sigma_u^2(1 - |u|^2)|b|^2 + \sigma_u^2\{I_j - |u|^2[2\mathrm{Re}[C(l)] - \hat{I}_j(l)]\} + N_o. \tag{23}$$

The conditional lth stage PSEP can be computed by the residue method as

$$\Pr(b \to \bar{b}|\mathbf{I}_\mathrm{j}) = \begin{cases} -\mathrm{Res}[\Phi_{\mathcal{D}(l)|\mathbf{I}_\mathrm{j}}(s)\frac{e^{s\delta}}{s}, p_2], & \text{if } \delta \le 0 \\ \mathrm{Res}[\Phi_{\mathcal{D}(l)|\mathbf{I}_\mathrm{j}}(s)\frac{e^{s\delta}}{s}, p_1] + \mathrm{Res}[\Phi_{\mathcal{D}(l)|\mathbf{I}_\mathrm{j}}(s)\frac{e^{s\delta}}{s}, 0], & \text{if } \delta > 0 \end{cases} \tag{24}$$

where $\mathrm{Res}[f, x]$ denotes the residue value of f at x, and $\delta = \delta(l)$.

B.1 PSEP of OD Scheme

For the DFM-OD scheme, we have $Y_1(l) = Z_1(l) = Y_1$, and, by (21) and (22), the conditional PSEP depends only on the MAI statistic I_j. By using (8), the UB and LB on the PSEP are given by

$$\mathrm{UB} = \sum_i \Pr\{b \to \bar{b} | I_j = \frac{i + K - 1}{N_u}\} d(i), \tag{25}$$

and

$$\mathrm{LB} = \sum_i \Pr\{b \to \bar{b} | I_j = \frac{i}{N_u}\} d(i). \tag{26}$$

The UB and LB on the APSEP can then be obtained via formulas (18), (25), and (26).

B.2 PSEP of SD Scheme

For the general DFM-SD scheme, by (21) and (22), the conditional PSEP depends on $Y_1(l)$ through $\hat{I}_j(l)$, and $Z_1(l)$ through $H_j(l)$, where $H_j(l) = I_j - |u|^2\{2\mathrm{Re}[C(l)] - \hat{I}_j(l)\}$. Due to the error propagation effect of the DFM detection, it is difficult to statistically characterize $H_j(l)$, and so far we are unable to derive the analytic results for the bounds on the PSEP of the general DFM-SD scheme. However, the conditional PSEP has been averaged over the channel fading, and Section B gives the formulations for the semi-analytic simulation of the PSEP of the general DFM-SD scheme. For the case where the 1st stage PSEP is considered, and the receiver has perfect knowledge of the past MAI, $I_j^{(k)}$, for $k = 2 \dots K$, the UB and LB on the PSEP can be obtained. We call the decision scheme the 1st SD scheme (SD1 scheme). The result serves as a lower bound on the PSEP of the 1st stage detection. In this case, $H_j(1)$ and $\hat{I}_j(1)$ are given by $H_j(1) = \sum_{k=2}^{K}[(1 - |u|^2)|I_j^{(k)}|^2 + |J_j^{(k)}|^2]$, and $\hat{I}_j(1) = \sum_{k=2}^{K} |I_j^{(k)}|^2$, respectively. By generalizing the results of Section III to the multidimensional case [8], the UB and LB on the PSEP of the DFM-SD1 scheme can be shown to be

$$\mathrm{UB} = \sum_{i_1, i_2} \Pr\{b \to \bar{b} | H_j(1) = \frac{i_1 + K - 1}{N_u}, \hat{I}_j(1) = \frac{i_2 + K - 1}{N_u}\} d(i_1, i_2), \tag{27}$$

and

$$\mathrm{LB} = \sum_{i_1, i_2} \Pr\{b \to \bar{b} | H_j(1) = \frac{i_1}{N_u}, \hat{I}_j(1) = \frac{i_2}{N_u}\} d(i_1, i_2), \tag{28}$$

where $d(\cdot, \cdot)$ is the pmf of the two-dimensional characterizing random vector for $H_j(1)$ and $\hat{I}_j(1)$.

C Effect of Estimation Error of u on PSEP

In this section, we assume that the correlation coefficient u defined in Section II is not perfectly known to the receiver, and investigate the resulting performance degradation. The results can be used to justify the use of the decision rules proposed in Section A for the mismatched Rayleigh fading channel. We assume that the DFM-OD scheme is used, and u is the true correlation coefficient. Suppose that the receiver uses u_1 instead of u in the decision metric (14). Then by generalizing the approach in Section B, we can show that, if $u_1^* u$ is real, the PSEP of the DFM-OD scheme is given by (24), where the corresponding terms are given by (20),

$$p_1 p_2 = -\frac{Y_2 \bar{Y}_2}{|u_1|^2 \sigma_u^2 |b - \bar{b}|^2 Z_2}, \tag{29}$$

$$w = \frac{(1 - |u_1|^2)(|b|^2 - |\bar{b}|^2) Z_3 + Y_2 u_1^* u |b - \bar{b}|^2 + (|u_1|^2 - u_1^* u)(Y_2 |\bar{b}|^2 - \bar{Y}_2 |b|^2)}{2|u_1|^2 |b - \bar{b}|^2 Z_2}, \tag{30}$$

and $\delta = 2\ln[Y_2/\bar{Y}_2]$. The parameters Y_2, Z_2 and Z_3 are defined as

$$Y_2 = \sigma_u^2(1 - |u_1|^2)|b|^2 + \sigma_u^2(1 - |u_1|^2) I_j + N_o, \tag{31}$$

$$Z_2 = \sigma_u^2(1-|u|^2)|b|^2 + \sigma_u^2(1-2u_1^*u+|u_1|^2)I_j + N_o, \tag{32}$$

and
$$Z_3 = \sigma_u^2(1-u_1^*u)|b|^2 + \sigma_u^2(1-2u_1^*u+|u_1|^2)I_j + N_o. \tag{33}$$

Note that the case with $u \neq u_1 = 1$ represents the system where the fading estimation is assumed to be exact, and the effect of mismatched fading is not considered in the receiver design.

V Conventional Single-User Detection

In this section, the decision rules and the PSEP are briefly reviewed for the following three CS detection schemes [8]: the optimum decision scheme (CS-OD), the 1st suboptimum decision scheme (CS-SD1), and the 2nd suboptimum decision scheme (CS-SD2). For the CS detection scheme, the noncentralized receiver is used. Since the 1st receiver does not utilize the decision feedback values or channel estimations $v_j^{(k)}$ for $k = 2 \ldots K$, the CSI vector of the 1st receiver for the CD scheme is given by $[v_j^{(1)}]$ only. Let the random vector $\mathbf{t}$ be defined as $\mathbf{t} = [r_j^{(1)}, v_j^{(1)}]^T$. Then by following the same approach as that in Section IV, the decision metrics and bounds on the PSEP can be derived for the three schemes, and are given in [8]. The results are summarized in the following. For the CS-OD scheme, the MAI statistic I_j is assumed to be known to the receiver; for the CS-SD1 scheme, I_j is approximated by its expected value M_I, which is as defined in (9); and for the CS-SD2 scheme, I_j is approximated by zero. It was shown in [8] that the pdf of I_j is highly centered around its expected value, and the performance of the CS-SD1 scheme is very close to that of the CS-OD scheme if the number of users K is known to the receiver. Also, the performance of the CS-SD2 scheme is close to that of the CS-OD scheme only when K is small.

VI Results

In this Section, we compute the UB and LB on the APSEP for the DFM-OD scheme, DFM-SD1 scheme, CS-OD scheme, CS-SD1 scheme, and CS-SD2 scheme. Gold sequences with length $N = 31$ [9], and rectangular chip waveforms are used. The parameter N_u, which controls the tightness of the bounds, equals 1000 throughout the examples. We compare the APSEP's of various schemes in the uncoded 16-QAM system for two error events. For event 1, $b = \frac{1}{\sqrt{10}}(1+i3)$, and $\bar{b} = \frac{1}{\sqrt{10}}(1+i)$. For event 2, $b = \frac{1}{\sqrt{10}}(1+i3)$, and $\bar{b} = \frac{1}{\sqrt{10}}(-3-i3)$. In Fig. 2, the mismatched fading with $u = 0.95$ is considered, and the signal to thermal noise ratio (STNR) E_b/N_o equals 20dB. We assume that u is known to the receiver. It is observed that the performance of the CS-SD1 scheme is very close to that of the CS-OD scheme for all values of K, and the performance of the CS-SD2 scheme is close to that of the CS-OD scheme only when K is small. Also, in some cases the performance of the CS-SD1 scheme is close to that of the DFM-SD1 scheme, which is a lower bound on the performance of the DFM-SD scheme at the 1st stage. Hence, if the complexity of the receiver is restricted to the one-stage detection, then the CS-SD1 scheme is as good as the one-stage DFM-SD scheme. Finally, the performances of the CS-SD1 and DFM-OD schemes serve as the upper and lower bounds on the performance of the DFM detection scheme; that is, the performance curve of the DFM-SD scheme at each stage lies between them. In Fig. 3, we assume that u is the true correlation coefficient, and the receiver uses u_1 instead of u in the decision metric for the DFM-OD scheme (see Section IV.C). Note that for the proposed DFM detection scheme, we have $u_1 = u$. The number of users K is 10 and only the event 1 is considered in this example. We see that the proposed system has some gain in STNR over the system where the estimation of the fading is assumed to be exact ($u_1 = 1$), and the results justify the use of the decision rules proposed in Section IV.A for the mismatched Rayleigh fading channel. Note that the error floor exists for those cases where $u < 0.999$, and becomes less obvious when $u > 0.9999$ in this example.

VII Conclusion

The most distinct features of this paper are as follows. We take into account the effects of mismatched Rayleigh fading on the design of the decision rules for the asynchronous DFM detection scheme. We obtain the tight UB and LB on the PSEP, which, unlike most of the other related papers, has been averaged over the channel fadings, time delays, and phase angles. And finally, compared with the time-consuming simulations for the performance evaluation of the CDMA systems, the analytic techniques proposed in this paper are much more efficient, especially when the number of users K is large, and provide exact results for some detection schemes of interest.

References

[1] S. Verdú, "Minimum probability of error for asynchronous Gaussian multiple access channels," *IEEE Trans. Inform. Theory.*, vol. IT-32, pp. 85-96, Jan. 1986.

[2] T. R. Giallorenzi and S. G. Wilson, "Decision-feedback multiuser receivers for asynchronous CDMA systems," in *Proc. GLOBECOM '93*, Houston, TX, 1993.

[3] T. R. Giallorenzi and S. G. Wilson, "Suboptimum multiuser receivers for convolutionally coded asynchronous DS-CDMA systems," *IEEE Trans. Commun.*, vol. COM-44, pp. 1183-1196, Sep. 1996.

[4] Y. C. Yoon and R. Kohno, "A spread-spectrum multiaccess system with cochannel interference cancellation for multipath fading channels," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 1067-1075, Sep. 1993.

[5] S. D. Gray, M. Kocic, and D. Brady, "Multiuser detection in mismatched multiple-access channels," *IEEE Trans. Commun.*, vol. COM-43, pp. 3080-3089, Dec. 1995.

[6] M. B. Pursley, "Performance evaluation for phase-coded spread-spectrum multiple-access communication - Part I: System analysis," *IEEE Trans. Commun.*, vol. COM-25, pp. 795-799, Aug. 1977.

[7] R. T. Hsu and J. S. Lehnert, "A characterization of multiple-access interference in generalized quadriphase spread-spectrum communications," *IEEE Trans. Commun.*, vol. COM-42, pp. 2001-2010, Feb. 1994.

[8] T.-T. Chen and J. S. Lehnert, "Bounds on the pairwise error probability of coded DS/SSMA communication systems in Rayleigh fading channels," to appear in *IEEE ICUPC 97'* ; also submitted to *IEEE Trans. Commun.*

[9] M. B. Pursley and H. F. A. Roefs, "Numerical evaluation of correlation parameters for optimal phases of binary shift-register sequences," *IEEE Trans. Commun.*, vol. COM-27, pp. 1597-1604, Oct. 1979.

[10] J. Cavers and P. Ho, "Analysis of the error performance of trellis-coded modulations in Rayleigh fading channels," *IEEE Trans. Commun.*, vol. COM-40, pp. 74-83, Jan. 1992.

[11] R. V. Nobelen and D. P. Taylor, "Analysis of the pairwise error probability of noninterleaved codes on the Rayleigh-fading channel," *IEEE Trans. Commun.*, vol. COM-44, pp. 456-463, Apr. 1996.

[12] M. Schwartz, W. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw-Hill, 1966.

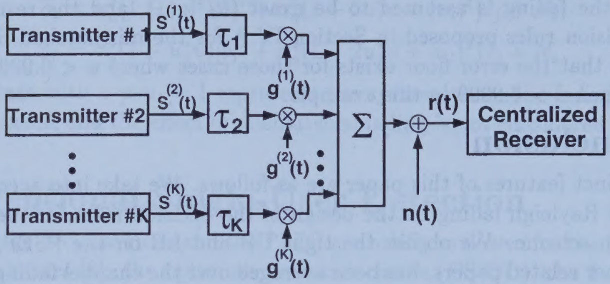

Transmitter #1
Transmitter #2
Transmitter #K
$s^{(1)}(t)$
$s^{(2)}(t)$
$s^{(K)}(t)$
τ_1
τ_2
τ_K
$g^{(1)}(t)$
$g^{(2)}(t)$
$g^{(K)}(t)$
Σ
r(t)
n(t)
Centralized Receiver

Figure 1: Baseband equivalent system model.

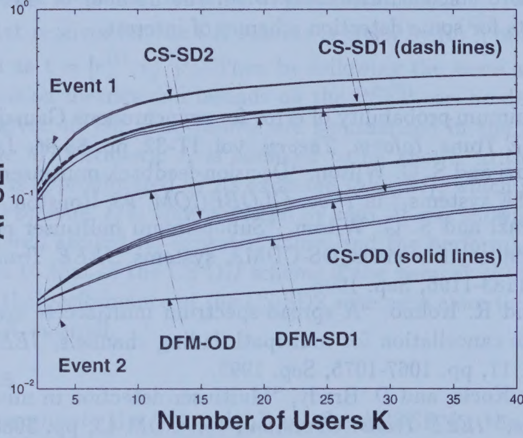


Figure 2: APSEP's of event 1 and event 2 for various decision schemes in the uncoded 16-QAM system with $u = 0.95$. The STNR E_b/N_o is 20dB.

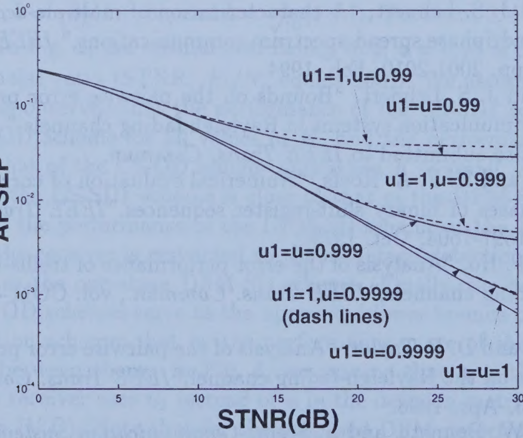


Figure 3: APSEP's of event 1 for the DFM-OD scheme in the uncoded 16-QAM system. The number of users K is 10.

Reduced Complexity Maximum Likelihood Receiver for Coded Multiuser Communication Systems

Moon-Hwan Yang and Maïté Brandt-Pearce
Department of Electrical Engineering
University of Virginia
Charlottesville, VA 22903

ABSTRACT

The optimal joint detector/decoder for a coded multiuser communication system becomes to complex to be implemented as the number of users gets large. In this paper it is shown that the trellis metric for this detector can be equivalently written as an affine transformation of trellis metrics each involving just one user. A reduced complexity detector/decoder is then developed by choosing the maxima of these single-user metrics and then linearly transforming these values to obtain the metric of the MUML receiver. The resulting multiuser detector shows excellent performance in exchange with a modest decrease in complexity.

1 Introduction

Multiuser communications systems employing such access schemes as code division multiple access (CDMA) require the use of either powerful error control coding or a multiuser detector to achieve high performance in the presence of many users and high additive Gaussian noise levels. The system considered in this paper uses both error control coding and multiuser detection, for which the computational complexity of the optimal joint decoding and detecting algorithm is prohibitively complex. The objective is to find powerful suboptimal decoder and multiuser detectors in a combined algorithm, called a *decoder/detector*.

The optimal multiuser detector is the multi-user maximum likelihood (MUML) receiver which it performs a maximum likelihood (ML) test on the entire set of users jointly rather than on each user individually. For uncoded systems, Verdú has shown that this algorithm can be performed using a trellis and has derived the metric for asynchronous additive white Gaussian noise (AWGN) channels [1]. For convolutionally coded links, Giallorenzi and Wilson have shown the structure of this receiver [2]. They define a super-code symbol that exists over a bit period and the crosscorrelations between the basis functions of this super-code symbol. Given these two terms and the energy of the users, they derive the metric of the receiver. They also show that the complexity of the receiver increases exponentially with the number of users and the constraint length of the convolutional code.

There are three major categories of suboptimal approaches to multi-user detection: trellis approaches, linear approaches, and decision feedback equalizers. For coded systems, one can either partition the decoding process from the multiuser detection process, or use a combined algorithm as proposed in this paper. All multiuser decoder/detector schemes previously studied for convolutionally encoded multiuser signals have been either partitioned approaches or decision feedback schemes (interference cancellation) [3, 4]. Trellis approaches are ones that directly stem from the MUML approach, by defining a sub-trellis which is likely to include the actual transmitted path. If the optimal metric is used and every symbol of every user is considered in the decoding process, MUML results but at a significant cost in complexity. However, considering a subgroup of users rather than the entire set of users allows a significant reduction in the complexity of the receiver, as shown by Varanasi with his group detection method for uncoded signals [5]. In this paper, we show that the MUML metric can be written as an affine combination of single-user metrics. We then propose a different trellis based suboptimal multiuser decoder/detector, named the reduced complexity MUML (RMUML) decoder/detector, where the MUML metric is obtained by choosing optimal single-user metrics and transforming them to obtain a smaller set of MUML metrics on which to base the multiuser decision.

2 Metric of the MUML Receiver

We begin by defining the received signal. In this paper, a synchronous transmission amongst the users is assumed, for notational simplicity. The extension of the algorithm to asynchronous systems is briefly discussed. The received signal before the matched filters can be expressed as the sum of N user's transmitted signals and additive white Gaussian noise, $n(t)$, with spectral density $N_0/2$,

$$r(t) \;=\; \sum_{i=-M}^{M} \sum_{k=1}^{N} \sqrt{E_k} D_k(t - iT_b) + n(t) \tag{1}$$

where T_b is the bit period, $2M+1$ is the size of the window, E_k is the bit energy of the kth user, and $D_k(t - iT_b)$ is the ith bit codeword waveform of the kth user. The received signal in (1) can be also be represented by a modulo N decomposition, namely with $i = \alpha(i)N + \beta(i) - 1$, as

$$r(t) = \sum_{i=-MN}^{MN} \sqrt{E_{\beta(i)}} D_{\beta(i)}(t - \alpha(i)T_b) + n(t) \tag{2}$$

For this decomposition, $\alpha(i)$ is the index of bits and $\beta(i)$ is the index of users.

For convolutionally coded systems in AWGN channels, assuming all the transmitted sequences are equally likely, the optimal decoder finds the sequence that maximizes the normalized likelihood function,

$$\tilde{D}^* = \arg\max_{\tilde{D}\in\mathcal{D}} e^{\Lambda(\tilde{D})} \tag{3}$$

where the set of codewords for each user is from a code with constraint length K, defined jointly as

$$\mathcal{D} = \prod_{k=1}^{N} \left\{S_{k,0}(t),\ S_{k,1}(t), ...,\ S_{k,2^K-2}(t),\ S_{k,2^K-1}(t)\right\}^{(2M+1)} \tag{4}$$

The codes for all users can be the same as long as some form of randomization is used to avoid very high crosscorrelations from occurring. The log-likelihood function can be expressed as

$$\begin{aligned} \Lambda(\tilde{D}) &= 2 \sum_{i=-MN}^{MN} \int_{-\infty}^{\infty} r(t) \left[\sqrt{E_{\beta(i)}} \tilde{D}_{\beta(i)}(t - \alpha(i)T_b)\right] dt \\ &\quad - \int_{-\infty}^{\infty} \left[\sum_{i=-MN}^{MN} \sqrt{E_{\beta(i)}} \tilde{D}_{\beta(i)}(t - \alpha(i)T_b) \right]^2 dt \\ &= \sum_{\alpha(i)} \lambda(\alpha(i)) \end{aligned} \tag{5}$$

where we have defined the single bit metric $\lambda(\alpha(i))$ as the sum of the metric over N users in the $\alpha(i)$th bit. To clarify the notation, let us define two quantities - the matched filter output for the $\alpha(i)$th bit of the $\beta(i)$th user as $\Gamma_{\beta(i)}(\alpha(i))$ and the energy of the sum of all the users' signals in the $\alpha(i)$th bit, $E(\alpha(i))$,

$$\Gamma_{\beta(i)}(\alpha(i)) = \int_{\alpha(i)T_b}^{(\alpha(i)+1)T_b} r(t) \left[\sqrt{E_{\beta(i)}} \tilde{D}_{\beta(i)}(t - \alpha(i)T_b)\right] dt$$

and

$$\begin{aligned} E(\alpha(i)) &= \int_{\alpha(i)T_b}^{(\alpha(i)+1)T_b} \left[\sum_{\beta(i)=1}^{N} \sqrt{E_{\beta(i)}} \tilde{D}_{\beta(i)}(t - \alpha(i)T_b) \right]^2 dt \\ &= \sum_{\beta(i)=1}^{N} E_{\beta(i)} + \sum_{\beta(i)=1}^{N} \sum_{\beta(j)=1, i \neq j}^{N} \int_{\alpha(i)T_b}^{(\alpha(i)+1)T_b} \left[\sqrt{E_{\beta(i)}} \tilde{D}_{\beta(i)}(t - \alpha(i)T_b) \right. \\ &\quad \left. \sqrt{E_{\beta(j)}} \tilde{D}_{\beta(j)}(t - \alpha(i)T_b)\right] dt. \end{aligned} \tag{6}$$

The single bit metric can now be expressed as

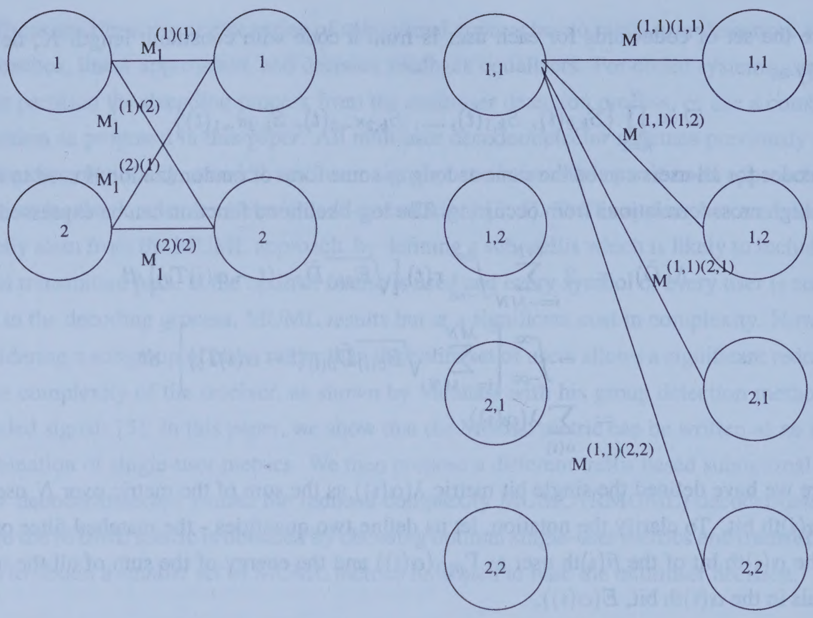
$$\lambda(\alpha(i)) = 2 \sum_{\beta(i)=1}^{N} \Gamma_{\beta(i)}(\alpha(i)) - E(\alpha(i)) \tag{7}$$

If a single user is active on the channel, the optimal decoder is a Viterbi algorithm which uses a metric referred to here as the single-user metric. This metric for the $\alpha(i)$th bit of the $\beta(i)$th user can be written as

$$\lambda_{\beta(i)}(\alpha(i)) = 2\Gamma_{\beta(i)}(\alpha(i)) - E_{\beta(i)}. \tag{8}$$

Using this notation, we can now express the single bit multiuser metric as

$$\begin{aligned} \lambda(\alpha(i)) &= \sum_{\beta(i)=1}^{N} \lambda_{\beta(i)}(\alpha(i)) - \sum_{\beta(i)=1}^{N} \sum_{\beta(j)=1, i \neq j}^{N} \int_{\alpha(i)T_b}^{(\alpha(i)+1)T_b} \left[\sqrt{E_{\beta(i)}} \tilde{D}_{\beta(i)}(t - \alpha(i)T_b) \right. \\ &\quad \left. \sqrt{E_{\beta(j)}} \tilde{D}_{\beta(j)}(t - \alpha(i)T_b)\right] dt \end{aligned} \tag{9}$$



(a) (b)

Figure 1: Trellises of (a) a single user and (b) a two user system for a $K = 2$ code. The number in each circle indicates the state and the expression on the lines indicate the metric for that transition.

Note that the MUML metric can be written as a linear transformation of single-user metrics plus a residual energy term. It is this form of the MUML metric that can be exploited in the development of a reduced complexity decoder/detector.

So far, we have considered the general expression of the MUML metric. Within the trellis, the code symbols $\tilde{D}$ can be determined by the state transitions of each user. For further clarification, we examine in detail a two user case as an example. The MUML metric in the $\alpha(i) = 0$ time interval can then be written

$$\begin{aligned}\lambda(0) &= \lambda_1(0) + \lambda_2(0) - 2\sqrt{E_1}\sqrt{E_2}\int_0^{T_b} \tilde{D}_1(t)\tilde{D}_2(t)\,dt \\ &= \lambda_1(0) + \lambda_2(0) - 2\sqrt{E_1}\sqrt{E_2}R(0) \qquad (10)\end{aligned}$$

where $R(0)$ is the residual crosscorrelation. For a two user system where each user is coded with a convolutional code with constraint length 2, each user has 2 states, labeled state 1 and 2. In order to obtain the states of a 2 user MUML receiver, the states of each user are combined for form four states. Now, the metric in (10) can be written as a state dependent metric, denoted as M. For the single-user trellis, the state dependent metric can be determined by choosing the proper codewords for the state transition, as shown in Figure 1, giving

$$M^{(1)(1)}_{\beta(i)} = \lambda_{\beta(i)}(0)|_{\tilde{D}_{\beta(i)}(t)=S_{\beta(i),0}(t),0<t\leq T_b}$$

$$\begin{aligned} M^{(2)(1)}_{\beta(i)} &= \lambda_{\beta(i)}(0)|_{\tilde{D}_{\beta(i)}(t)=S_{\beta i,1}(t),\,0<t\leq T_b} \\ M^{(1)(2)}_{\beta(i)} &= \lambda_{\beta(i)}(0)|_{\tilde{D}_{\beta(i)}(t)=S_{\beta i,2}(t),\,0<t\leq T_b} \\ M^{(2)(2)}_{\beta(i)} &= \lambda_{\beta(i)}(0)|_{\tilde{D}_{\beta(i)}(t)=S_{\beta(i),3}(t),\,0<t\leq T_b}, \end{aligned} \tag{11}$$

for $\beta(i) = 1, 2$, where the superscript in M is the state transition and the subscript is the index of the user.

For the two-user MUML trellis, the state dependent metric M can be obtained by choosing the codewords for each single-user metric and the residual energy term $R(0)$. For example,

$$\begin{aligned} M^{(1,1)(1,1)} &= M_1^{(1)(1)} + M_2^{(1)(1)} - 2\sqrt{E_1}\sqrt{E_2}R(0)|_{\tilde{D}_1(t)=S_{1,0}(t),\tilde{D}_2(t)=S_{2,0}(t),\,0<t\leq T_b} \\ M^{(1,1)(1,2)} &= M_1^{(1)(1)} + M_2^{(1)(2)} - 2\sqrt{E_1}\sqrt{E_2}R(0)|_{\tilde{D}_1(t)=S_{1,0}(t),\tilde{D}_2(t)=S_{2,2}(t),\,0<t\leq T_b} \\ M^{(1,1)(2,1)} &= M_1^{(1)(2)} + M_2^{(1)(1)} - 2\sqrt{E_1}\sqrt{E_2}R(0)|_{\tilde{D}_1(t)=S_{1,2}(t),\tilde{D}_2(t)=S_{2,0}(t),\,0<t\leq T_b} \\ M^{(1,1)(2,2)} &= M_1^{(1)(2)} + M_2^{(1)(2)} - 2\sqrt{E_1}\sqrt{E_2}R(0)|_{\tilde{D}_1(t)=S_{1,2}(t),\tilde{D}_2(t)=S_{2,2}(t),\,0<t\leq T_b}. \end{aligned} \tag{12}$$

In our notation, the first parenthesis in the superscript is the departure states and the second parenthesis is the destination state. For example, (1,2) means the first user is at state 1 and the second user is at state 2. A two-user MUML metric for the $\alpha(i) = 0$ bit can be presented in matrix form, emphasizing its structure as an affine transformation of single-user metrics, which we write as

$$\left(\begin{array}{c} M^{(1,1)(1,1)} \\ M^{(1,2)(1,1)} \\ M^{(2,1)(1,1)} \\ M^{(2,2)(1,1)} \\ \hline M^{(1,1)(1,2)} \\ M^{(1,2)(1,2)} \\ M^{(2,1)(1,2)} \\ M^{(2,2)(1,2)} \\ \hline M^{(1,1)(2,1)} \\ M^{(1,2)(2,1)} \\ M^{(2,1)(2,1)} \\ M^{(2,2)(2,1)} \\ \hline M^{(1,1)(2,2)} \\ M^{(1,2)(2,2)} \\ M^{(2,1)(2,2)} \\ M^{(2,2)(2,2)} \\ \hline \end{array}\right) = \left(\begin{array}{c} 10001000 \\ 10000100 \\ 01001000 \\ 01000100 \\ 10000010 \\ 10000001 \\ 01000010 \\ 01000001 \\ 00101000 \\ 00100100 \\ 00011000 \\ 00010100 \\ 00100010 \\ 00100001 \\ 00010010 \\ 00010001 \end{array}\right) \left(\begin{array}{c} M_1^{(1)(1)} \\ M_1^{(2)(1)} \\ \hline M_1^{(1)(2)} \\ M_1^{(2)(2)} \\ \hline M_2^{(1)(1)} \\ M_2^{(2)(1)} \\ \hline M_2^{(1)(2)} \\ M_2^{(2)(2)} \\ \hline \end{array}\right) - 2\sqrt{E_1}\sqrt{E_2} \left(\begin{array}{c} R^{(1,1)(1,1)} \\ R^{(1,2)(1,1)} \\ R^{(2,1)(1,1)} \\ R^{(2,2)(1,1)} \\ \hline R^{(1,1)(1,2)} \\ R^{(1,2)(1,2)} \\ R^{(2,1)(1,2)} \\ R^{(2,2)(1,2)} \\ \hline R^{(1,1)(2,1)} \\ R^{(1,2)(2,1)} \\ R^{(2,1)(2,1)} \\ R^{(2,2)(2,1)} \\ \hline R^{(1,1)(2,2)} \\ R^{(1,2)(2,2)} \\ R^{(2,1)(2,2)} \\ R^{(2,2)(2,2)} \\ \hline \end{array}\right) \tag{13}$$

where $R^{(i_1,i_2)(j_1,j_2)}$ is the crosscorrelation between the two user symbols which cause a transition between state (i_1, i_2) and state (j_1, j_2). In (13), metrics with the same destination are grouped together. In a Viterbi decoder, we choose the optimal path to each destination state and then starting from the end of the trellis connect the surviving paths to obtain the optimal overall path. It is therefore more convenient to group the metrics with the same destination as in (13).

3 Metric of the RMUML Receiver

In Section 2, the MUML metric is represented as a linear transformation of single-user's metrics plus the residual energy term. In this section, we propose a trellis-based sub-optimum detector named the reduced complexity MUML (RMUML) decoder/detector that uses this structure of the MUML metric. The RMUML detector uses single-user trellis decisions to form a reduced number of MUML metrics for comparison. In the two user case, if two single-user receivers are used, then we can choose a maximum from each of the four blocks in the single-user vector. Having done so and then linearly transforming this result, we can recover four MUML metrics, one from each of four blocks in the MUML metric vector. The RMUML detector then makes a multiuser symbol decision based on this reduced number of hypotheses. For MUML detection, we would have had to calculate 16 metrics. Since the maxima of single-users' metrics are transformed to obtain the MUML metric, the maximum likelihood path will often be preserved.

For example, if the following single-user metrics are chosen

$$\begin{aligned} M_1^{(1)(1)} > M_1^{(2)(1)},\ M_1^{(1)(2)} > M_1^{(2)(2)}, \\ M_2^{(1)(1)} > M_2^{(2)(1)},\ M_2^{(1)(2)} > M_2^{(2)(2)} \end{aligned} \tag{14}$$

then, (13) reduces to

$$\begin{pmatrix} \underline{M^{(1,1)(1,1)}} \\ \underline{M^{(1,1)(1,2)}} \\ \underline{M^{(1,1)(2,1)}} \\ \underline{M^{(1,1)(2,2)}} \end{pmatrix} = \begin{pmatrix} 1010 \\ 1001 \\ 0110 \\ 0101 \end{pmatrix} \begin{pmatrix} \underline{M_1^{(1)(1)}} \\ \underline{M_1^{(1)(2)}} \\ \underline{M_2^{(1)(1)}} \\ \underline{M_2^{(1)(2)}} \end{pmatrix} - 2\sqrt{E_1}\sqrt{E_2} \begin{pmatrix} \underline{R^{(1,1)(1,1)}} \\ \underline{R^{(1,1)(1,2)}} \\ \underline{R^{(1,1)(2,1)}} \\ \underline{R^{(1,1)(2,2)}} \end{pmatrix} \tag{15}$$

Note that the reduced matrix identifying the transformation from the single-user metric vector to the multiuser metric vector is dependent of the actual single-user decoder decisions made. This does not pose an otherwise expected implementation difficulty since the multiuser metric vector is simply the sum of the surviving paths for each state of each user, minus the corresponding crosscorrelation term. The actual reduced matrix need not be identified.

For this small example, the suboptimal decoder/detector method provides a reduction in computation complexity by a factor of four in trade with degraded performance. In general, given N users and each user uses an orthogonal convolutional code with constraint length K on a binary information signal, we have $2^{N(K-1)}$ states and 2^{NK} metrics to compute in a MUML receiver (2^N metrics per state). However, with the help of our proposed algorithm, only $N2^{K-1}$ single-user metrics are stored and by linearly transforming these single-user metrics, $2^{N(K-1)}$ MUML metrics can be recovered. The computational complexity is a factor of $2^{N(K-1)}/2^{NK} = 1/2^N$ smaller. Although the complexity of the RMUML receiver is still exponentially dependent on N, the reduction in complexity is also exponentially dependent on N and, for K typically small, there is a significant reduction in complexity. Table 1 compares the number of trellis metrics using both methods for typical values of N and K. It is clear

System	MUML	RMUML
2 user, K = 4	256	64
2 user, K = 8	65536	16384
4 user, K = 4	65536	4096
4 user, K = 8	4.2950e+9	2.68435e+8
10 user , K = 4	1.0995e+12	1.0737e+9

Table 1: Complexity comparison between the MUML and the RMUML receivers.

that for a small system with short codes, the RMUML decoder/detector provides a substantial computational advantage. For a system with large N and K, neither detector is feasible.

4 Simulation Results

We simulate the performance of a simple two user system to determine the degradation in performance from reducing the number of hypotheses in the MUML algorithm. To get a reasonable performance level, the constraint lengths for both users are chosen to be 4. The codes used are orthogonal convolutional codes and the truncation depth is chosen to be 100. In Figures 2 and 3, the bit error probability of user 1 and user 2 are plotted, respectively. E_1/N_0 is fixed at 4 dB while E_2/N_0 varies from 1 dB to 7 dB. In both plots, the single-user performance bound and the single-user receiver are used as references for comparison. The single-user performance bound is the optimal performance when only one user is present along with background noise. The single-user receiver only considers the matched filter output of the single user in a Viterbi decoder configured for that user alone. All the multiuser receivers should perform somewhere between these two bounds. As seen from the plots, the average bit error probability for this example is very high since a constraint length of only $K = 4$ is used and no spreading is used. For Figure 2, the single-user performance bound is constant since the SNR of user 1 does not change. The MUML and the RMUML detectors show very consistent performance throughout the whole SNR region. The RMUML receiver performs better than the single-user receiver, yet does not appear to be near-far resistant [6], which is expected since it is based on single-user detector decisions.

We have considered the bit synchronous case so far. For the asynchronous case, both the MUML and the RMUML receivers operate in essentially the same ways as for the synchronous case. The values of the metric would change due to the relative delays between users but the complexity of the receiver would not change.

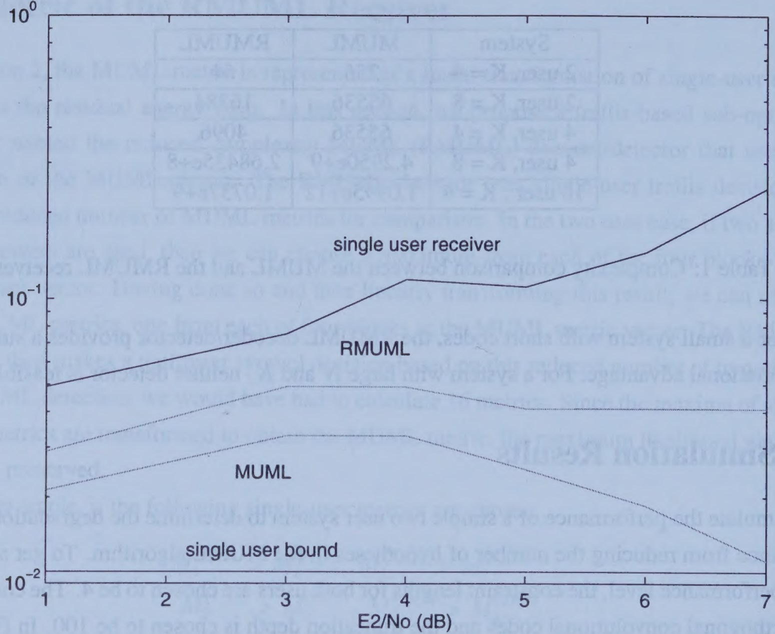


Figure 2: Simulation of bit error probability of user 1 for a two user system where E_1/N_0 is 4 dB and E_2/N_0 varies from 1 dB to 7 dB. Both users have constraint length $K = 4$.

5 Conclusion

In convolutionally coded CDMA, the computational complexity of the optimal MUML receiver is prohibitively high. Therefore, a suboptimal approach with lower complexity and a reasonable degradation in performance is introduced. The metric of the MUML receiver can be represented as an affine transformation of single-user metrics and the computation complexity can be significantly reduced by pre-choosing maxima of each single-user's metrics and then linearly transforming them to obtain the MUML metrics. The resulting RMUML detector reduces the complexity of the receiver by a factor equal to the exponent of the number of users and still provides acceptable performance, especially when compared to the conventional single-user detector.

References

[1] S. Verdú, "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Transactions on Communications*, vol. IT-32, pp. 85–96, January 1996.

[2] T. R. Giallorenzi and S. G. Wilson, "Multiuser ML sequence estimator for convolutionally coded asynchronous DS-CDMA systems," *IEEE Transactions on Communications*,

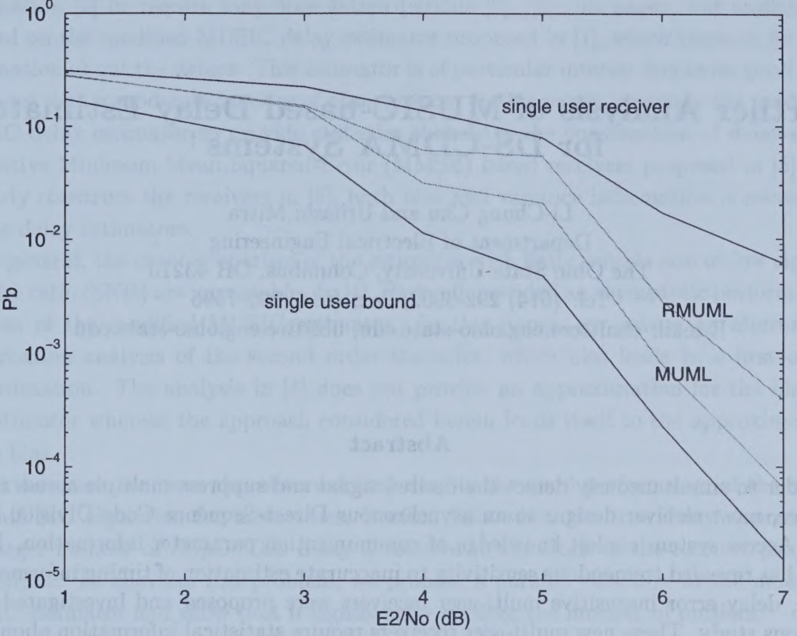


Figure 3: Simulation of bit error probability of user 2 of a two user system where E_1/N_0 is 4 dB and E_2/N_0 varies from 1 dB to 7 dB. Both users have constraint length $K = 4$.

vol. 44, pp. 997 – 1008, August 1996.

[3] Y. Sanada and M. Nakagawa, "A multiuser interference cancellation technique utilizing convolutional codes and orthogonal multicarrier modulation for wireless indoor communications," *IEEE J. Selected Areas Commun.*, vol. 14, no. 8, pp. 1500–1509, October 1996.

[4] A. Hafeez and W. E. Stark, "Combined decision-feedback multiuser detection/ soft-decision decoding for CDMA channels," *Vehicular Technology Society 45th Conference, Mobile Communication for Human Race*, pp. 382–386, Atlanta, GA, April 1996.

[5] M. K. Varanasi, "Group detection for synchronous Gaussian code-division multiple-access channels," *IEEE Transactions on Information Theory*, vol. 41, pp. 1083–1096, July 1995.

[6] S. Verdú, "Optimum multiuser asymptotic efficiency," *IEEE Trans. Commun.*, vol. COM-34, no. 9, pp. 890–897, Sept, 1986.

Further Analysis of MUSIC-based Delay Estimators for DS-CDMA Systems [†]

Li-Chung Chu and Urbashi Mitra
Department of Electrical Engineering
The Ohio State University, Columbus, OH 43210
Tel: (614) 292-3902, Fax: (614) 292-7596
E-mail: chul@er4.eng.ohio-state.edu; ubli@ee.eng.ohio-state.edu

Abstract

In order to simultaneously detect the desired signal and suppress multiple access interference, most receiver designs in an asynchronous Direct-Sequence Code-Division Multiple Access system exploit knowledge of communication parameter information. Prior work has revealed tremendous sensitivity to inaccurate estimation of timing information. Thus, delay error insensitive multi-user receivers were proposed and investigated in a previous study. These new multi-user receivers require statistical information about delay estimates in order to be constructed. Thus the current work focuses on a statistical analysis of a previously proposed MUSIC algorithm which requires no initial information and no exhaustive search of the delays. The current study is alternative and further to that previously performed for the modified MUSIC algorithm. In addition, a new arrangement of the MUSIC estimators is proposed to reduce the number of spurious estimates which are very far from the true delay, which is a major problem of the original algorithm.

I. INTRODUCTION

A significant number of the previously proposed receiver designs for multi-user receivers in Direct-Sequence Code-Division Multiple-Access (DS-CDMA) environments assume that any necessary communication parameters were perfectly known. This parameter knowledge was exploited to successfully suppress multiple access interference (MAI) as well as detect the signal of the user of interest. Of the various relevant parameter estimation problems in the multi-user context, acquisition of the signals is one of the most challenging. Thus, there has been considerable effort devoted to the development of delay estimators for DS-CDMA systems. The conventional sliding correlator suffers in a near-far scenario when the signal power of the interfering users overwhelms that of the desired user [1]. Other delay estimation algorithms either assume initial

---

[†] This research was supported by the Air Force Office of Scientific Research under Grant F49620-96-1-0105.

information [2] or require long observation periods [3]. In this paper, our analysis is focused on the modified MUSIC delay estimator proposed in [1], which requires no side information about the delays. This estimator is of particular interest due to its good performance with a moderate number of observations. In this paper, we study the modified MUSIC delay estimator to provide statistics that drive the construction of delay-error insensitive Minimum-Mean Squared-Error (MMSE) based receivers proposed in [5]. To properly construct the receivers in [5], both bias and variance information is necessary for the delay estimators.

In general, the exact properties of the estimates with finite sample size or low signal-to-noise ratio (SNR) are intractable. In [4], Parkvall provided an asymptotic performance analysis of the modified MUSIC estimator. In this paper, we pursue an alternative perturbation analysis of the second order statistics, which also leads to a first order approximation. The analysis in [4] does not provide an approximation for the bias of the estimator whereas the approach considered herein lends itself to the approximation of the bias.

Furthermore, we have also observed that despite the good performance of the modified MUSIC algorithm for most cases, it often fails when the true delay approaches an integer number of chips. This delay is relative to the clock of the receiver prior to acquisition. To alleviate this problem, we propose a parallel structure of the modified MUSIC estimator and show that it significantly reduces the number of outliers.

This paper is organized as follows. In Section II, the system model and assumptions are provided. In addition, the modified MUSIC algorithm is reviewed. In Section III, the new performance analysis of the modified MUSIC algorithm is provided. The bias and variance estimates are compared to previous estimates and simulation results. The notion of the twin MUSIC estimator is introduced in Section IV and numerical results are also presented. Final conclusions are found in Section V.

II. SYSTEM MODEL AND MODIFIED MUSIC ALGORITHM

The system model and the MUSIC algorithm summarized in this section are adopted from [1]. In a K-user asynchronous DS-CDMA system, each user employs a spreading waveform $b_k(t)$ with unit energy to form the baseband signal. These spreading waveforms are formed by modulating rectangular chip waveforms with binary pseudo-noise sequences. The total received signal is the sum of signals from all active users plus white Gaussian noise. Under the assumption of coherent detection, this baseband signal can be represented as:

$$r(t) = \sum_{k=1}^{K} \sum_{m=-\infty}^{\infty} d_k(m) b_k(t - mT) + n(t), \tag{1}$$

where the data $d_k(m) \in \{+1, -1\}$, for BPSK modulation. It is assumed that the data $d_k(m)$ are identically distributed and independent for all k (users) and m (time). Furthermore, the data and the noise processes are also independent. The received signal is

IQ-mixed and then matched filtered with the chip pulse shape. After sampling at the chip-rate, the m^{th} frame of the received signal from the kth user can be expressed as

$$\mathbf{r}_k(m) = [\mathbf{a}_{2k-1}\ \ \mathbf{a}_{2k}] \begin{bmatrix} \beta_k & 0 \\ 0 & \beta_k \end{bmatrix} \begin{bmatrix} z_{2k-1}(m) \\ z_{2k}(m) \end{bmatrix}, \tag{2}$$

where the received amplitude is $\beta_k = \sqrt{P_k} e^{j\theta_k}$. The data frames are of length N, which correspond to the number of chips in the spreading waveform. The vectors $\mathbf{a}_{2k-1}$ and $\mathbf{a}_{2k}$ represent the shifted signal sequences modulated by different combinations of two adjacent binary symbols, z_{2k-1} or z_{2k},

$$\begin{aligned} z_{2k-1}(m) &= \frac{1}{2}(d_k(m) + d_k(m-1)), \\ z_{2k}(m) &= \frac{1}{2}(d_k(m) - d_k(m-1)). \end{aligned} \tag{3}$$

The shifted code sequences are defined as

$$\mathbf{a}_{2k-1} = \left[\delta_k \mathbf{D}_{+1}^{(p_k+1)} + (1-\delta_k)\mathbf{D}_{+1}^{(p_k)}\right] \mathbf{c}_k \tag{4}$$

$$\mathbf{a}_{2k} = \left[\delta_k \mathbf{D}_{-1}^{(p_k+1)} + (1-\delta_k)\mathbf{D}_{-1}^{(p_k)}\right] \mathbf{c}_k \tag{5}$$

where p_k and δ_k denote the integer and decimal part of the delay, respectively. Thus $\tau_k = p_k T_c + \delta_k$. The length N vector $\mathbf{c}_k$ is the spreading vector of user k. The permutation matrix $\mathbf{D}_s^{(p)}$ is defined in block form as,

$$\mathbf{D}_s^{(p)} = \begin{bmatrix} 0 & \mathbf{I}_{n-p} \\ s\mathbf{I}_p & 0 \end{bmatrix} \tag{6}$$

If $N > 2K$, the autocorrelation matrix $\mathbf{R} = \mathbf{E}\{\mathbf{r}(m)\mathbf{r}(m)^*\}$ is rank deficient and can thus be decomposed in signal and noise subspaces. The signal subspace will have dimension $2K$. In practice, the autocorrelation matrix must be estimated. An unbiased estimate of the correlation matrix and its eigenvalue decomposition (EVD) are given by

$$\hat{\mathbf{R}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{r}(m)\mathbf{r}^*(m) = \hat{\mathbf{E}}_s \hat{\mathbf{\Lambda}}_s \hat{\mathbf{E}}_s^* + \hat{\mathbf{E}}_n \hat{\mathbf{\Lambda}}_n \hat{\mathbf{E}}_n^*, \tag{7}$$

where M is the number of data frames. The subscript s denotes quantities relating to the signal subspace and the subscript n denotes quantities relating to the noise subspace. Thus $\hat{\mathbf{E}}_s$ and $\hat{\mathbf{E}}_n$ span the estimated signal and noise subspaces, respectively.

The user of interest is labeled as user 1 and the subscript $k = 1$ is dropped. The basic idea behind the MUSIC algorithm is to determine the delay value that minimizes the projection of the signal space defined by Span$\{\mathbf{a}_{2k-1}, \mathbf{a}_{2k}\}$ onto the noise space. According to the MUSIC algorithm, the delay estimate $\hat{\tau}$, is then the argument that minimizes the cost function J,

$$\hat{\tau} = \arg\min J(\tau) = \arg\min \left\{\mathbf{a}^*(\tau)\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*\mathbf{a}(\tau)\right\}, \tag{8}$$

where $\mathbf{a}(\tau)$ is the observed signal component corresponding to a relative delay of τ between the signal arrival and the starting edge of the observation frame. The minimization of this cost function is problematic in the context of CDMA signals due to the nature of the spreading waveforms employed. Recall that these waveforms are composed of rectangular chip pulses. Thus, $J(\tau)$ is analytic only when τ varies inside one chip interval.

A feasible approach to combat the lack of analyticity is to separate the estimation process into N sub-estimation problems [1]. In each of the problems, the delay estimate is constrained to be within one chip interval, where the cost function is analytic, so the MUSIC algorithm may be applied directly. As asynchronous transmission is considered, the received signal from the first user is a linear combination of two vectors, $\mathbf{a}_1$ and $\mathbf{a}_2$. Hence, the generalized cost function for the q^{th} sub-estimator is a weighted squared norm of the 2-D manifold spanned by $\mathbf{a}_1$ and $\mathbf{a}_2$ onto the the noise space $\hat{\mathbf{E}}_n$,

$$J_q(\tau_q) = \mathrm{Tr}\{\mathbf{U}_q\mathbf{A}_q^*(\tau_q)\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*\mathbf{A}_q(\tau_q)\}\Big|_{\tau_q\in[(q-1)T_c,qT_c]}, \tag{9}$$

where the weighting matrix $\mathbf{U}_q$ is a 2 by 2 real matrix, and $\mathbf{A}_q$ is an N by 2 matrix with column vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ corresponding to the delay in the q^{th} chip interval. $\mathrm{Tr}\{\cdot\}$ is the trace operator. The minimum of J_q is a consistent estimate if the true delay τ resides in $((q-1)T_c, qT_c)$ [6].

The resulting overall estimate is given by

$$\begin{aligned} \hat{\tau} &= \arg\min_{\hat{\tau}_1\cdots\hat{\tau}_N} J_q(\hat{\tau}_q) \\ \text{where} \quad \hat{\tau}_q &= \arg\min J_q(\tau_q) \end{aligned} \tag{10}$$

Thus, the estimate is selected from one of the N sub-estimates. Due to the constraint that the sub-estimates must reside in the corresponding chip interval, there is no guarantee that a correct selection will be made for a finite sample size. We shall refer to the spurious estimates corresponding to the incorrect selection as "outliers". That is, these outliers correspond to the incorrect chip interval altogether. Since the statistical properties of the outliers are intractable, the analytical performance provided in Section III excludes the outliers as done in [1, 4]. However, we shall investigate the behavior of the outliers via simulations in Section IV.

III. Performance Analysis of the MUSIC Estimator

In this section, we present a performance analysis of the modified MUSIC estimator. Our objective is develop estimates for the first and second moments of the estimation error. This analysis is alternative and further to that considered in [4]. The approach is similar to [7], which strictly follows a perturbation technique based on a Taylor series expansion. The weighting matrix $\mathbf{U}_q$ is assumed to be constant for all q (*i.e.* $\mathbf{U}_q = \mathbf{U}\ \forall\, q$), to simplify the expressions of the partial derivatives. It is straightforward to modify the derivation for general $\mathbf{U}_q$ matrices.

We assume the perturbation is small and find the Taylor series expansion of $\dot{J}_q$ as a function of the delay τ_q and the eigenvectors of the signal space, which are denoted $\mathbf{e}_i, i = 1, \ldots, 2K$,

$$\dot{J}_q\left(\hat{\tau}_q, \hat{\mathbf{E}}_s\right) \approx \dot{J}_q + \ddot{J}\Delta\tau_q + 2\mathrm{Re}\sum_{i=1}^{2K} \mathbf{g}_i^T \Delta\mathbf{e}_i. \tag{11}$$

The vector $\mathbf{g}_i$ is defined to be the directional gradient of $\dot{J}_q$ along $\mathbf{e}_i$,

$$\mathbf{g}_i = \nabla_{\mathbf{e}_i}\dot{J} = -\mathrm{Tr}\{\mathbf{U}\mathbf{A}^c\Psi^T + \mathbf{U}^T\Psi^c\mathbf{A}^T\}\mathbf{e}_i, \tag{12}$$

where $\mathbf{A} = [\mathbf{a}_1\mathbf{a}_2]$ and $\Psi = \frac{\partial A}{\partial \tau}$. The superscript 'c' denotes the conjugate, 'T' denotes the transpose, and '*' denotes the conjugate transpose. All of the terms on the right-hand side of (11) are evaluated at the nominal point $(\tau_q, \mathbf{E}_s)$. The perturbation of the delay that minimizes the perturbed cost function is obtained by setting (11) to zero, giving

$$\Delta\tau_q \approx \frac{-2\mathrm{Re}\sum_{i=1}^{2K} \mathbf{g}_i^T \Delta\mathbf{e}_i}{\ddot{J}_q}, \tag{13}$$

where the fact that $\dot{J}_q(\tau_q, \mathbf{E}_s) = 0$ has been applied. The second moment of $\Delta\tau_q$ can then be obtained by standard manipulation of (13),

$$E[\Delta\tau_q^2] \approx \frac{2}{\ddot{J}_q^2}\mathrm{Re}\sum_{i=1}^{2K}\sum_{j=1}^{2K}\left\{\mathbf{g}_i^T E[\Delta\mathbf{e}_i\Delta\mathbf{e}_j^T]\mathbf{g}_j + \mathbf{g}_i^T E[\Delta\mathbf{e}_i\Delta\mathbf{e}_j^*]\mathbf{g}_j^c.\right\} \tag{14}$$

where the perturbed eigenvectors are functions of the eigenvalues (λ_i) and eigenvectors $(\mathbf{e}_i)$ of $\mathbf{R}$ [8],

$$E\left[\Delta\mathbf{e}_i\Delta\mathbf{e}_j^*\right] \approx \delta_{ij}\frac{\lambda_i}{M}\sum_{\substack{k=1\\k\neq i}}^{N}\frac{\lambda_k}{(\lambda_i - \lambda_k)^2}\mathbf{e}_k\mathbf{e}_k^*; \tag{15}$$

$$E\left[\Delta\mathbf{e}_i\Delta\mathbf{e}_j^T\right] \approx (1 - \delta_{ij})\frac{-\lambda_i\lambda_j}{M(\lambda_i - \lambda_j)^2}\mathbf{e}_j\mathbf{e}_i^T. \tag{16}$$

The function δ_{ij} is the Kronecker delta function and is zero when $i \neq j$ and unity otherwise.

Bias of the *modified* MUSIC delay estimator of [1] has not been previously addressed in the literature; though bias analysis has been considered for the regular MUSIC estimator [9] as it is an important performance characteristic for limited sample size or low SNR. In addition, such information is necessary for the construction of the delay error insensitive estimators of [5]. To fully utilize the statistical properties of the eigenvectors up to the second order, we expand $\dot{J}_q$ by a second-order Taylor series expansion,

$$\begin{aligned}\dot{J}_q\left(\hat{\tau}_q, \hat{\mathbf{E}}_s\right) \approx\ & \dot{J}_q + \ddot{J}_q\Delta\tau_q + 2\mathrm{Re}\sum_{i=1}^{2K}\mathbf{g}_i^T\Delta\mathbf{e}_i + \frac{1}{2}\,\dddot{J}_q\,\Delta\tau_q^2 + 2\mathrm{Re}\sum_{i=1}^{2K}\left[\mathbf{q}_i^T\Delta\mathbf{e}_i\Delta\tau_q\right] \quad (17)\\ & + \mathrm{Re}\ \mathrm{Tr}\left[\sum_{i,j=1}^{2K}\mathbf{H}_{i,j}\Delta\mathbf{e}_j\Delta\mathbf{e}_i^* + \breve{\mathbf{H}}_{i,j}\Delta\mathbf{e}_j\Delta\mathbf{e}_i^T\right],\end{aligned}$$

where the vector $\mathbf{q}_i$ is the directional gradient of $\ddot{J}_q$ along $\mathbf{e}_i$,

$$\mathbf{q}_i = \nabla_{\mathbf{e}_i} \ddot{J} = -\mathrm{Tr}\{(\mathbf{U} + \mathbf{U}^T)\mathbf{\Psi}^c\mathbf{\Psi}^T\}\mathbf{e}_i. \tag{18}$$

The Hessian matrices can be expressed as

$$\begin{aligned} [\mathbf{H}_{i,j}]_{k,l} &= \frac{\partial^2 \dot{J}}{\partial \mathbf{e}_i^c(k)\partial \mathbf{e}_j(l)} \qquad (19) \\ &= \begin{cases} \left[\frac{1}{2}\mathrm{Tr}\{\mathbf{U}\mathbf{A}^c\mathbf{\Psi}^T + \mathbf{U}^T\mathbf{\Psi}^c\mathbf{A}^T\}\right]_{k,l}, & i = j \\ 0, & i \neq j \end{cases} \\ \left[\breve{\mathbf{H}}_{i,j}\right]_{k,l} &= \frac{\partial^2 \dot{J}}{\partial \mathbf{e}_i(k)\partial \mathbf{e}_j(l)} = 0. \qquad (20) \end{aligned}$$

It can be shown that $\ddot{J}(\tau_q, \mathbf{E}_s)$ is not a function of τ_q, thus, $\dddot{J}_q\,(\tau_q, \mathbf{E}_s) = 0$. Therefore, the expectation of $\Delta\tau_q$ that satisfies $\dot{J}(\hat{\tau}_q, \hat{\mathbf{E}}_s) = 0$ can be approximated by

$$E[\Delta\tau_q] \;\approx\; \frac{-1}{\ddot{J}_q}\mathrm{Re}\sum_{i=1}^{2K}\left\{2\mathbf{g}_i^T E[\Delta\mathbf{e}_i] + 2\mathbf{q}_i^T E[\Delta\mathbf{e}_i\Delta\tau_q] + \mathrm{Tr}\{\mathbf{H}_{i,i}E[\Delta\mathbf{e}_i\Delta\mathbf{e}_i^*]\}\right\}, \tag{21}$$

where the bias of the eigenvectors can be obtained from the EVD of $\mathbf{R}$ [8],

$$E[\Delta\mathbf{e}_i] = \frac{-\lambda_i}{2M}\sum_{\substack{k=1 \\ k\neq i}}^{N}\frac{\lambda_k}{(\lambda_i - \lambda_k)^2}\mathbf{e}_i. \tag{22}$$

The approximation of $E[\Delta\mathbf{e}_i\Delta\tau_q]$ is given by premultiplying (13) with $\Delta\mathbf{e}_i$ before taking the expectation [9],

$$E[\Delta\mathbf{e}_i\Delta\tau_q] \approx \frac{-1}{\ddot{J}_q}\sum_{j=1}^{2K}\left\{E[\Delta\mathbf{e}_i\Delta\mathbf{e}_j^*]\mathbf{g}_j^c + E[\Delta\mathbf{e}_i\Delta\mathbf{e}_j^T]\mathbf{g}_j\right\}. \tag{23}$$

Finally, the bias is obtained by substituting (15), (22) and (23) into (21).

Figure 1 shows the analytical and simulated results of the MUSIC estimator with 5 and 10 users. The symbols of each user are spread by Gold codes of length N=31. The signal-to-noise ratio, defined as symbol energy over the two-sided power spectral density of noise $N_0/2$, is 15 dB. Each interfering user is 20 dB stronger than the desired user. We assume that $\mathbf{U} = \begin{bmatrix} 1/|\mathbf{a}_1|^2 & 0 \\ 0 & 1/|\mathbf{a}_2|^2 \end{bmatrix}$ is a fixed and known matrix for the analytical, the approximated, and the simulated results. The simulation results are drawn from 500 Monte-Carlo runs. It is observed that the bias estimate overestimates the true bias, while the variance estimators tend to underestimate the true value. The approximation of MSE by (14) is slightly better than that in [1]. The accuracy of the approximation depends on the convergence rate of the Taylor series expansion to the true $J(\tau, \mathbf{E}_s)$. This convergence has not been fully studied at this time.

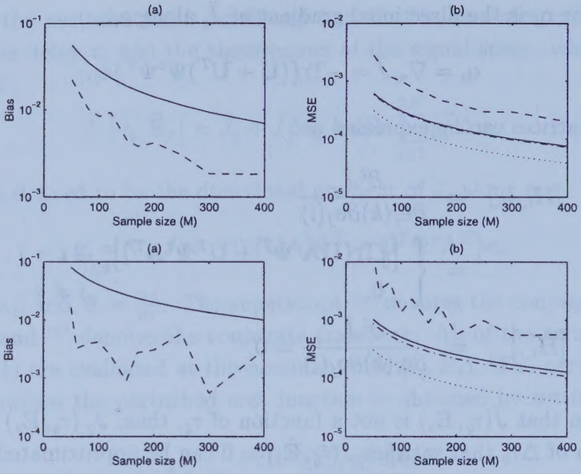


Figure 1: Approximation and simulation of estimator performance: Figures (a) and (b) are for 5 users; (c) and (d) are for 10 users. Dashed line is simulation; solid line is approximation by (14) or (21); dashed-dotted line and dotted line are approximation of MSE and asymptotic Cramèr-Rao bound by Ström *et al.* [1], respectively.

IV. TWIN MUSIC DELAY ESTIMATORS

Due to the fact that the cost function of the MUSIC algorithm discussed in Section II is not analytic at the chip boundary, the probability of outliers converges to zero with a very slow rate when the true delay is close to an integer number of chips. This problem can be resolved by applying a pair of the MUSIC estimators whose inputs have a relative delay of $0.5T_c$, as shown in Figure 2. The estimate with the lower cost is selected from the twin estimates. Since the boundary of the chip interval perceived by the second set of sub-estimators has a $0.5T_c$ delay with respect to the first set, the boundary effect will be compensated for amongst the two sets of sub-estimators. While pathological combinations of spreading codes and delays exist for which the probability of outliers for the twin estimators can be higher than that of the single version, this scenario has not been observed for the simulations conducted. In fact, it has been observed that the majority of outliers for the twin estimators tend to be concentrated near the true value with the means of these outlier clusters being less than $0.5T_c$ away from the true delay. Such an error is acceptable for many applications, (*e.g.*, [10]).

Table 1 shows the percentage of outliers for the single and twin estimators for two realizations of the delay: $\tau = 10.67T_c$ and $\tau = 10.03T_c$. As noted earlier, the second scenario poses a problem for the single estimator. For these numerical results, M=100 samples were employed. There were ten users in the system, utilizing length 31 Gold sequences. The SNR was 15db and the near-far ratio was 20 dB. The single estimates not in $[10T_c, 11T_c$ and the twin estimates not in $[9.5T_c, 11T_c]$ are counted as outliers. While it is not surprising that the twin estimator outperforms the single version for $\tau = 10.03T_c$,

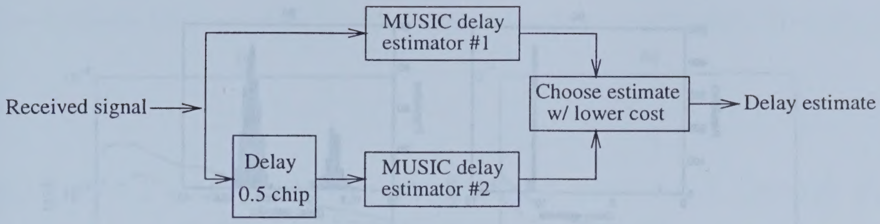


Figure 2: Block diagram of the twin MUSIC estimator.

M	40	60	80	100	150	200
single ($\tau = 10.67T_c$)	79.40	48.80	31.40	12.40	2.40	1.40
twin ($\tau = 10.67T_c$)	41.40	12.60	4.40	0.60	0.00	0.00
single ($\tau = 10.03T_c$)	70.00	49.20	47.80	44.60	39.20	34.00
twin ($\tau = 10.03T_c$)	20.20	2.00	0.20	0.00	0.20	0.00

Table 1: Percentage of outliers for single and twin MUSIC estimators.

the advantage also appears for short samples sizes when $\tau = 10.67T_c$. However, there is a performance cost incurred from the use of the twin estimator. This is observed in the histograms of the delay estimates ($\hat{\tau}$) shown in Figure 3. The system parameters are the same as those cited for Table 1. Figure 3 (a) depicts the histogram for the single estimator for $\tau = 10.67T_c$, similarly, Figure 3 (b) displays the histogram for the twin estimator. It is clear that there a finite probability that the incorrect member of the pair of the twin estimators will have the lowest cost. Figures 3 (c) and (d) show the same set of histograms for the scenario where $\tau = 10.03T_c$. Thus for delays that are not close to integer multiples of a chip, the mean-squared error (MSE) for twin estimates will be higher. This is seen in Figure 4 which provides the MSE for the estimators excluding the outliers. Figure 4 (a) provides the MSE plots for the two estimators for $\tau = 10.67T_c$. The solid line is for the twin estimator, while the dashed line is for the single estimator. As predicted, the twin estimator's MSE is slightly higher than that of the single estimator. However, in Figure 4 (b) where the solid line is for the single estimator and the dashed line is for the twin estimator, the twin estimator offers improved MSE. Thus the twin estimator trades off variance for a reduction in catastrophic outliers.

V. CONCLUSIONS

In this paper, we have provided further analysis of the statistics of the modified MUSIC estimator [1]; in particular, a bias estimate is developed. The bias and variance statistics are necessary for the development of improved MMSE receivers which have greater insensitivity to errors in timing information [5]. Numerical simulations

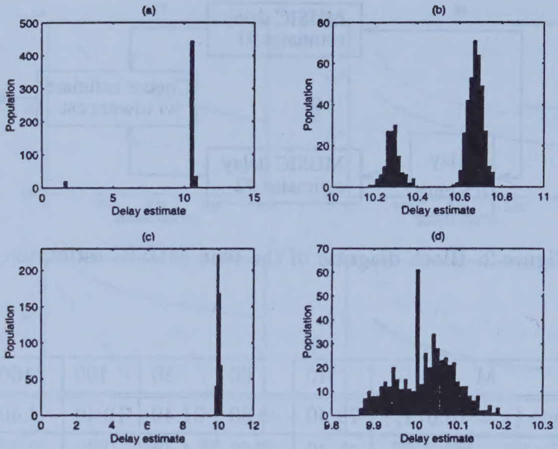


Figure 3: Performance of single and twin sub-estimators. The system parameters: $M = 100$; $N = 31$; 10 users; $\text{SNR}_1 = 15$ dB; NFR = 20 dB; 500 realizations. Figures (a) and (b) are the histograms of delay estimates by the single and the twin estimators, respectively for $\tau_1 = \mathbf{10.67}T_c$. Figures (c) and (d) are the histograms of delay estimates by the single and the twin estimators, respectively for $\tau_1 = \mathbf{10.03}T_c$.

show that the new moment estimates are slightly better than those previously derived, although both sets of estimator statistics are derived using perturbation analyses. Furthermore, tt had been observed that the modified MUSIC estimator often failed when the the true delay was close to an integer multiple of chips. This is due to the fact that the modified MUSIC algorithm's cost function is not analytic at chip boundaries. To alleviate this problem, a twin estimator structure is proposed and studied via simulation. The parallel structure of the twin estimator reduces the number of outliers and often offers overall improved MSE performance versus the original single estimator. Current work is focusing on quantifying the improvements in performance due to the use of the twin estimator in conjunction with the improved MMSE receiver of [5]. In addition, the sensitivity of the receiver in [5] to incorrect estimates of the input statistics is also under consideration.

REFERENCES

[1] E. Ström, S. Parkvall, S. L. Miller, and B. E. Ottersten, "Propagation delay estimation in asynchronous direct-sequence code-division multiple access systems," *IEEE Trans. on Communications*, vol. 44, pp. 84–93, January 1996.

[2] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1553–1560, October 1988.

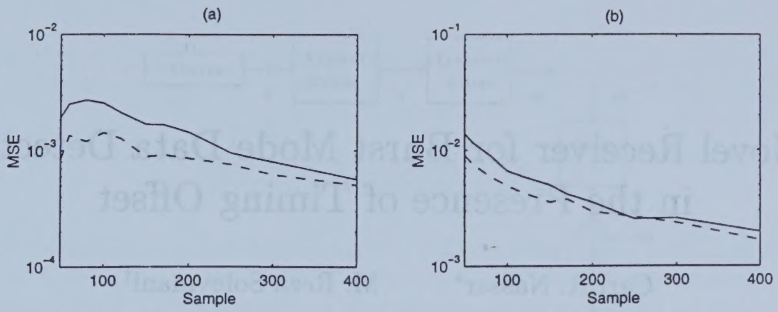


Figure 4: MSE of the single and twin sub-estimators as a function of the number of samples excluding outliers. Communication parameters: 10 users; $\text{SNR}_1 = 15$ dB; NFR $= 20$ dB; 500 realizations. Figure (a) shows the MSE for $\tau_1 = \mathbf{10.67}T_c$. The solid line is for the twin estimator and the dashed line is for the single estimator. Figure (b) shows the MSE for $\tau_1 = \mathbf{10.03}T_c$. The solid line is for the single estimator and the dashed line is for the twin estimator.

[3] R. A. Iltis and L. Mailaender, "Multiuser code acquisition using parallel decorrelators," in *Proc. of the Conference on Information Sciences and Systems*, pp. 109–114, Princeton, NJ, March 1994.

[4] S. Parkvall, *Near-Far Resistant DS-CDMA Systems: Parameter Estimation and Data Detection.* PhD thesis, Royal Institute of Technology, 1996.

[5] L.-C. Chu and U. Mitra, "Improved MMSE-based multi-user detectors for mismatched delay channels," in *Proc. of Conference on Information Sciences and Systems*, vol. 1, pp. 326–331, Princeton, NJ, March 1996.

[6] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramèr-Rao bound," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 720–741, May 1989.

[7] R. Hamza and K. Buckley, "An analysis of weighted eigenspace methods in the presence of sensor errors," *IEEE Trans. on Signal Processing*, vol. 43, pp. 1140–1150, May 1995.

[8] M. Kaveh and A. J. Barabell, "The statistical performance of the MUSIC and the minimum-norm algorithms in resolving plane waves in noise," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, pp. 331–341, April 1986.

[9] X.-L. Xu and K. M. Buckley, "Bias analysis of the MUSIC location estimator," *IEEE Trans. on Signal Processing*, vol. 40, pp. 2559–2569, October 1992.

[10] R. A. Iltis and L. Mailaender, "Multiuser detection of quasi-synchronous CDMA signals using linear decorrelators," *IEEE Trans. on Communications*, vol. 44, pp. 1561–1571, November 1996.

A Novel Receiver for Burst Mode Data Detection in the Presence of Timing Offset

Carl R. Nassar* M. Reza Soleymani†

Abstract

This paper introduces a novel receiver for the detection of data in the presence of noise and timing offset. The timing offset is assumed constant over each block of N symbols, but changes from one block of N symbols to the next (modeling the timing offset in burst mode communications). The receiver is derived by starting from joint data detection and timing offset estimation, and then introducing a novel quantization of the timing offset space. When compared to receivers available to date, our receiver demonstrates substantial performance gains in a number of important practical situations. Furthermore, the complexity of our receiver is very low.

1 Introduction

In many communication environments of practical interest, independent data symbols are sent over a channel introducing two effects. First, the channel adds a noise, characterized as white and Gaussian, and, second, it introduces a timing offset. In the case of burst mode transmission, which has found widespread usage in satellite and mobile communications, the timing offset introduced by the channel can often be described as changing from burst to burst.

Traditionally, receivers detect data in the presence of noise and timing offset using a simple approach. Tracking circuitry, either feedback or feedforward, is used to track and remove the timing offset; then, the data is detected in the presence of the noise alone using well-known Bayesian methods. Some of the analog tracking circuits can be found in [1][2].

In cases of burst mode transmission, the timing offset tracking circuits that are most suitable are the feedforward methods. One feedforward tracking circuit, presented by Oerder and Meyer [3], has recently gained widespread popularity. The popularity of this circuit is due in large part to its digital nature – the received signal is sampled by a free running oscillator and all the timing offset recovary is then done digitally. However, this tracking method has its drawbacks. First, as the roll-off factor α decreases, the

---

*Department of Electrical Engineering, Colorado State University, Ft Collins, Colorado 80523-1373

†Department of Electrical Engineering, Concordia University, Montreal, Quebec, Canada

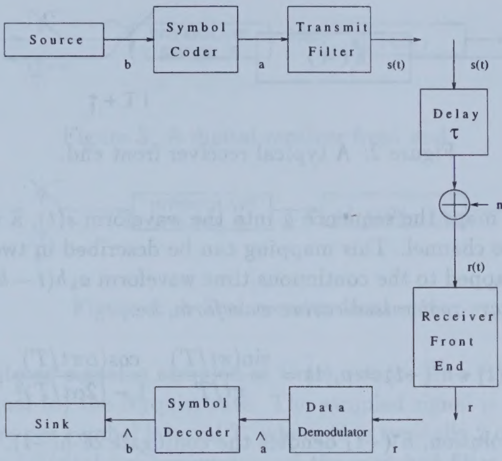


Figure 1: The communication system model.

performance of the tracking circuit diminishes; in fact, when α is close to 0, a receiver using this tracking circuit is unable to achieve reliable data detection. Another drawback of this tracking circuit is that its performance diminishes substantially as burst length decreases.

In what follows, we introduce a novel receiver, intended for burst mode transmission, that detects data in the presence of timing offset and noise. Unlike the receivers to date, this receiver does not explicitly track and remove the timing offset. Instead, its implementation corresponds to joint data detection and timing estimation. This receiver demonstrates substantial performance gains when compared to a receiver employing the timing recovary scheme in [3]. These gains are most significant in the important practical cases of small roll-off factors (α), as well as in cases of short burst length. Furthermore, our receiver demonstrates a low complexity; its complexity matches that of a receiver with perfect knowledge of timing offset over all but the first 10 symbols of the burst.

2 Communication System Model

This section introduces the model of the timing-offset communication environment of interest in this paper. With this in hand, we can, in future sections, present our novel receiver.

The model for the timing-offset communication environment is shown in Figure 1. Here, a *source* creates a sequence of binary digits labeled $\underline{b} = (b_1, ..., b_V)$. A *symbol coder* then maps these bits into the sequence of symbols $\underline{a} = (a_1, ..., a_N)$, where N is the burst length. These symbols are generated (sequentially) at a rate of one symbol every T second interval. Furthermore, each a_i is independent of a_j for all $j \neq i$; and, each a_i corresponds to a value in $A = \{a^1, ..., a^M\}$ — an example of an $a^K \in A$ is $a^K = e^{j\frac{2\pi}{M}K}$.

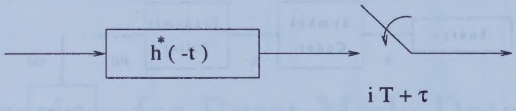


Figure 2: A typical receiver front end.

The *transmit filter* maps the sequence $\underline{a}$ into the waveform $s(t)$, a waveform ready for transmission over the channel. This mapping can be described in two parts. First, each data symbol a_k is mapped to the continuous time waveform $a_k h(t-kT)e^{jw_c t}$, where $h(t)$ corresponds to a *square-root raised-cosine waveform*, i.e.,

$$h(t) * h^*(-t) = p_s(t) = \frac{\sin(\pi t/T)}{\pi t/T} \cdot \frac{cos(\alpha \pi t/T)}{1-(2\alpha t/T)^2}\,; \tag{1}$$

here, $*$ denotes convolution, $h^*(-t)$ denotes the conjugate of $h(-t)$, and α is a value in [0,1] called the *roll-off factor*. The waveform $s(t)$ is then created by adding together all the $a_k h(t-kT)e^{jw_c t}$ terms, i.e.,

$$s(t) = \sum_{k=1}^{N} a_k h(t-kT)e^{jw_c t}. \tag{2}$$

(In practice, $s(t)$ is simply the real component of equation (2). However, we use complex notation here because it simplifies the presentation without impacting the receiver design.)

The *channel* demonstrates two effects: it introduces a timing offset, and it adds a noise. The output of the channel is described by

$$r(t) = s(t-\tau) + \eta(t). \tag{3}$$

Here, $\eta(t)$ represents the additive noise, modeled as AWGN (additive white Gaussian noise). Additionally, τ represents the timing offset, a constant over the N symbol burst. This offset represents a portion of the symbol duration, i.e., $0 \leq \tau < T$.

The *receiver front end* maps the waveform $r(t)$ into the sequence $\underline{r} = (r_1, r_2, ..., r_N)$, a sufficient statistic for detection. It does this using three components: a mixer, a matched filter, and a sampler. The mixer translates the symbol to baseband; we assume that the mixer has perfect knowledge of the phase and frequency. The output of this mixer is described by

$$r'(t) = \sum_{k=1}^{N} a_k h(t-\tau-kT) + \eta'(t), \tag{4}$$

where $\eta'(t)$ represents the noise after the mixer operation. The remainder of the receiver front end is shown in Figure 2. Here, the matched filter demonstrates an impulse response of $h^*(-t)$, and the sampler generates one symbol every T second interval.

In the case at hand, we employ a digital version of this receiver front end. Dropping the mixer for simplicity in presentation, a digital receiver front end is shown in Figure 3. A low pass filter removes the noise components outside of the information-bearing signal's

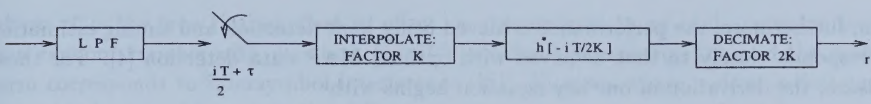


Figure 3: A digital receiver front end.

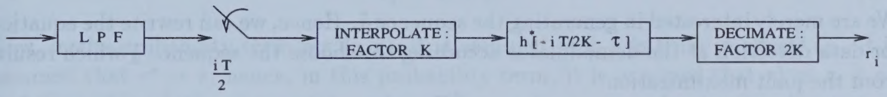


Figure 4: A final receiver front end

bandwidth. This filtered signal is sampled at time instants $\frac{iT}{2}+\tau$, a sampling at a rate greater than (or equal to) the Nyquist rate. The sampled signal is then interpolated to achieve $2\cdot K$ samples per symbol interval T, where K is typically 2 or 3. This upsampled signal is filtered by a digital implementation of the matched filter. Lastly, the filtered signal is downsampled by a factor of $2K$, leaving only the samples at time instants $iT+\tau$.

The digital receiver front end we employ corresponds to an alternative (equivalent) implementation of the digital receiver front end described by Figure 3. This alternative implementation is shown in Figure 4 (without the mixer). Here, the sampler using the correct sampling delay of τ has been replaced by a free running sampler with delay τ' (which, for simplicity in presentation, but without loss of generality, we assume is 0). The timing offset is now accounted for by an appropriate shift of the matched filter taps.

We present the receiver front end of interest in the case at hand by making a final realization. The value of τ is not available for the matched filter of Figure 4. Consequently, this matched filter is not easily realized. We will include the design of the matched filter (along with the interpolator and decimator) in the demodulator. Hence, the receiver front end we consider consists simply of a mixer, a low pass filter, and a sampler generating samples at time instants $\frac{iT}{2}$. The $2\times N$ long sequence output by this updated receiver front end, also a sufficient statistic for detection, is labeled $\underline{r}' = (r'_{1,1}, r'_{1,2}, r'_{2,1}, r'_{2,2}, ..., r'_{N,1}, r'_{N,2})$.

Returning to describe the components in Figure 1, the *demodulator* implements the discrete time matched filter (along with the interpolator and decimator), outputting the sequence $\underline{r}$; it then maps $\underline{r}$ into $\underline{\hat{a}}$, an estimate of the transmitted symbol sequence. A simple *symbol decoder* follows, mapping the symbols $\underline{\hat{a}}$ into bits $\underline{\hat{b}}$, an estimate of the source's binary sequence.

3 The Novel Receiver

3.1 Key Equation

In this section, we introduce an equation characterizing our novel demodulator. We generate this equation by using, as our starting point, the jointly optimal data detection and timing estimation equation. Joint detection and timing estimation is chosen for two reasons: first, it is more easily implemented than the optimal MAP data detection equation; furthermore, the performance achieved using joint detection and timing estimation corresponds closely to that achieved with optimal MAP data detection [4]. For these reasons, the derivation of our key equation begins with

$$\hat{\underline{a}}, \hat{\tau} = \arg \max_{\underline{a},\tau} p(\underline{a}, \tau | \underline{r}'). \tag{5}$$

We are merely interested in generating the sequence $\hat{\underline{a}}$. Hence, we can rewrite the equation for data detection at the demodulator according to: choose the sequence $\hat{\underline{a}}$ which results from the joint maximization

$$\max_{\underline{a},\tau} p(\underline{a}, \tau | \underline{r}'). \tag{6}$$

This equation can be simplified by applying statistical and mathematical operations, and by making a few simple observations. This leads to

$$\max_{\underline{a},\tau} p(\underline{r}' | \underline{a}, \tau) \cdot p(\tau). \tag{7}$$

We can further simplify this equation by using the available information regarding τ. This information is the following. The value τ is a constant, unknown value over the block of N symbols. This τ can be characterized statistically according to:

$$p(\tau) = \begin{cases} \frac{1}{T}, & \tau \in \Upsilon = [0, T) \\ 0, & \text{else} \end{cases} . \tag{8}$$

Applying this τ information to equation (7) leads to:

$$\max_{\underline{a},\tau} p(\underline{r}' | \underline{a}, \tau). \tag{9}$$

Because logarithms are monotonic functions, this maximization can be rewritten as

$$\max_{\underline{a},\tau} \ln\, p(\underline{r}' | \underline{a}, \tau). \tag{10}$$

Reordering the maximizations leads to

$$\max_{\tau}[\max_{\underline{a}} \ln\, p(\underline{r}' | \underline{a}, \tau)]. \tag{11}$$

The probability in the maximization, $p(\underline{r}' | \underline{a}, \tau)$, can be restated in a more convenient form. Specifically, this probability corresponds to $p(\underline{r}^\tau | \underline{a}, \tau)$, where $\underline{r}^\tau = (r_1^\tau, r_2^\tau, ..., r_N^\tau)$ refers to the output of the matched filter in the front end of the demodulator, when it receives $\underline{r}'$ and assumes τ is the timing offset in effect. Using this probability in equation (11) results in

$$\max_{\tau}[\max_{\underline{a}} \ln\, p(\underline{r}^\tau | \underline{a}, \tau)]. \tag{12}$$

We can further simplify our equation by examining the probability $p(\underline{r}^\tau | \underline{a}, \tau)$, the probability of $\underline{r}^\tau$ when $\underline{a}$ and τ are assumed to be in effect. First, we observe that the values in the sequence $\underline{r}^\tau$ (the sequence output by a matched filter given $\underline{r}'$ and assuming τ) correspond to

$$r_i^\tau \simeq a_i p_s(\tau^* - \tau) + \sum_{k \neq i} a_k p_s((i-k)T + (\tau^* - \tau)) + \eta_i, \tag{13}$$

where τ^* is the true timing offset in effect, η_i corresponds to an i.i.d. complex Gaussian random variable with variance $\frac{N_o}{2}$, $p_s(\cdot)$ is defined in equation (1), and the middle term corresponds to intersymbol interference (ISI). In cases where τ does in fact truly correspond to τ^*, this output simplifies to

$$r_i^\tau \simeq a_i + \eta_i. \tag{14}$$

Now, in the probability term $p(\underline{r}^\tau|\underline{a},\tau)$, it is assumed that τ is truly in effect, i.e., it is assumed that $\tau^* = \tau$; hence, in this probability term, it is assumed that $r_i^\tau \simeq a_i + \eta_i$. It follows that $p(\underline{r}^\tau|\underline{a},\tau) = p_{\underline{\eta}}(\underline{r}^\tau - \underline{a}) = \prod_{i=1}^N p_{\eta_i}(r_i^\tau - a_i)$, or, in shorthand notation, $p(\underline{r}^\tau|\underline{a},\tau) = \prod_{i=1}^N p(r_i^\tau|a_i)$. Introducing this equality to equation (12) results in

$$\max_{\tau} \sum_{i=1}^{N} [\max_{a_i} \ln\, p(r_i^\tau|a_i)]. \tag{15}$$

The final equation, which sets up the novel demodulator implementation, is achieved by introducing an approximation to equation (15). The continuous timing offset space $\Upsilon = [0,T)$ is approximated by a discrete space $\tilde{\Upsilon} = \{\tau^1,...,\tau^m\}$, $\tilde{\Upsilon} \subset \Upsilon$. This approximation is a very good one, because we insure that data detection achieved using $\tilde{\Upsilon}$ is very close to that achieved using Υ. The elements of the space $\tilde{\Upsilon}$ are the topic of Section 3.3. For now, we use the generic $\tilde{\Upsilon} = \{\tau^1,...,\tau^m\}$ and return to deriving the key equation. Applying the approximation of Υ by $\tilde{\Upsilon}$ to equation (15) leads to: choose the sequence $\underline{\hat{a}}$ from the joint maximization

$$\max_{\tilde{\tau}\in\tilde{\Upsilon}} [\sum_{i=1}^{N} \max_{a_i} \ln p(r_i^{\tilde{\tau}}|a_i)]. \tag{16}$$

3.2 Implementation

In this subsection, we provide an implementation for our demodulator. This implementation corresponds to a parallel evaluation of equation (16).

The demodulator implementation is shown in Figure 5. There are two main parts to this implementation. The m parallel branches carry out the inner maximization of equation (16), maximizing over a_i. The *computation and decision unit (CDU)*, meanwhile, carries out the outer maximization of equation (16), maximizing over $\tilde{\tau}$.

The m parallel branches work as follows. In the i^{th} symbol interval $[iT,(i+1)T)$, the j^{th} branch receives $\underline{r}'(i)$, the subset of the sequence $\underline{r}'$ required to generate the matched filter output at the i^{th} symbol time. This j^{th} branch assumes that $\tau = \tau^j$. With $\underline{r}'(i)$ and the assumed value $\tau = \tau^j$ in hand, the j^{th} branch first applies a *matched filter* demonstrating the impulse response $h[\frac{-iT}{2K} - \tau^j]$. The output of this matched filter is labeled r_i^j. A *decision device* is then applied at the j^{th} branch. As it also assumes $\tau = \tau^j$, the decision device believes the received sample r_i^j corresponds to $r_i^j = a_i + \eta_i$. Consequently, it generates the ML decision

$$\hat{a}_i^j = \arg\max_{a_i} p(r_i^j|a_i) = \arg\max_{a_i} p_{\eta_i}(r_i^j - a_i). \tag{17}$$

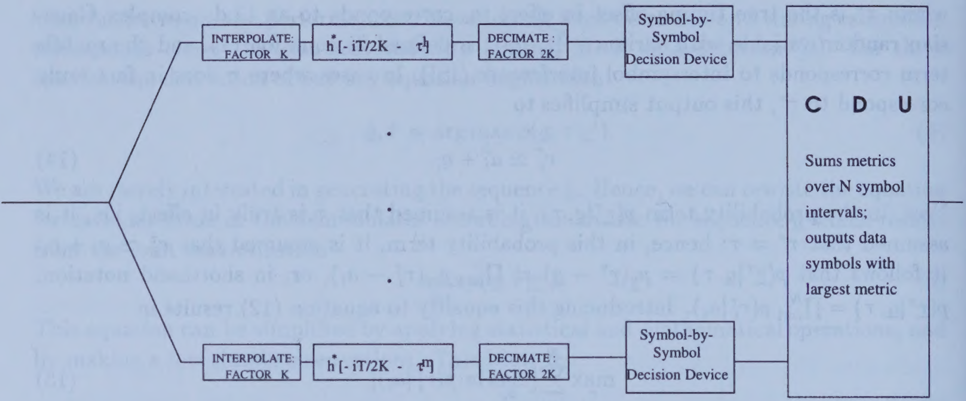

INTERPOLATE: FACTOR K
$h^*[- iT/2K - \tau^1]$
DECIMATE : FACTOR 2K
Symbol-by-Symbol Decision Device
C D U
Sums metrics over N symbol intervals; outputs data symbols with largest metric
INTERPOLATE: FACTOR K
$h^*[- iT/2K - \tau^m]$
DECIMATE : FACTOR 2K
Symbol-by-Symbol Decision Device

Figure 5: The demodulator implementation.

It also generates the corresponding likelihood value

$$l_i^j = \max_{a_i} p(r_i^j|a_i) = \max_{a_i} p_{\eta_i}(r_i^j - a_i) = p_{\eta_i}(r_i^j - \hat{a}_i^j). \tag{18}$$

The processing at the j^{th} branch corresponds to the implementation of the inner maximization of equation (16), namely $max_{a_i} \ln p(r_i^{\tilde{\tau}}|a_i)$, when evaluated at $\tilde{\tau} = \tau^j$; the logarithm in the maximization has not been included for convenience.

The *CDU* generates the N-symbol output $\underline{\hat{a}}$ by selecting a set of decisions from among the many decisions of the m parallel branches. Specifically, it selects its decisions by implementing the outer maximization of equation (16). The CDU output, based on equation (16), can be expressed in terms of the parallel branches' decisions according to

$$\underline{\hat{a}} = (\hat{a}_1^{j^*}, ..., \hat{a}_N^{j^*}),\ j^* \in \{1, 2, ..., m\}; \tag{19}$$

here, j^* corresponds to the index of the branch whose decisions are selected by the CDU. This value is generated according to

$$j^* = \arg\max_j \sum_{i=1}^{N} \ln l_i^j. \tag{20}$$

The CDU implementing the above two equations is described as follows. The CDU sums the N likelihoods it receives from each branch; it then determines which sum is the largest; finally, it outputs all of the decisions of the branch with the largest sum.

3.3 The Discrete Space $\tilde{\Upsilon}$

In this subsection, we present the discrete space $\tilde{\Upsilon} = \{\tau^1, ..., \tau^m\}$ for use in the key equation and corresponding implementation. The discrete space $\tilde{\Upsilon}$ is generated in two parts. First, a starting size m is established using an algorithm based on rate distortion theory. Second, the exact size and elements of $\tilde{\Upsilon}$ are generated using an algorithm based

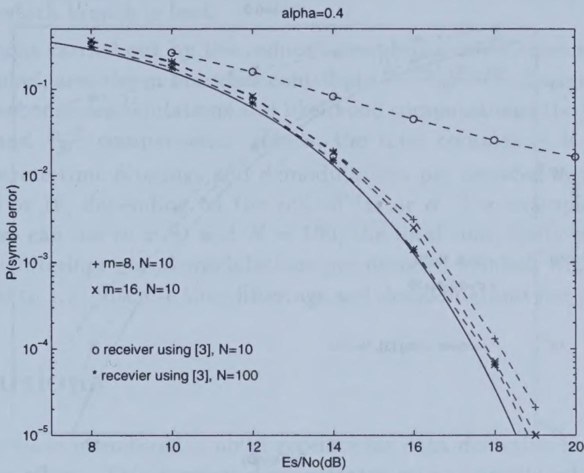


Figure 6: $P(\epsilon) - \frac{E_s}{N_o}$ for 8-PSK, $\alpha = 0.4$.

on the Generalized Lloyd Algorithm. The details of this analysis can be found in [5][6]. After some computation, we find that a discrete space $\tilde{\Upsilon} = \{\tau^1, ..., \tau^m\}$ that achieves a detection performance very close to that achieved using $\Upsilon = [0, T)$ is the following: $\tilde{\Upsilon}$ corresponds to a uniformly spaced set of m values between $[0, T)$, i.e.,

$$\tilde{\Upsilon} = \{\frac{1}{2m}T, \frac{3}{2m}T, ..., \frac{2m-1}{2m}T\}, \tag{21}$$

where $m = 8$ or $m = 16$ depending on the roll-off factor. Roll-off factors very close to 0 require $m = 16$; otherwise, $m = 8$ suffices.

4 Performance and Complexity

4.1 Performance

The performance of our novel receiver is generated by computer simulation, using a_i's corresponding to 8-PSK symbols.

Figures 6 and 7 present the performance of our receiver using $P(\epsilon)$ vs $\frac{E_s}{N_o}$ plots. These figures present the performance of our receiver along with both coherent performance and the performance of a receiver using the popular timing tracking of [3].

Figure 6 shows the performance results when the raised cosine waveform $p_s(t)$ demonstrates a roll-off of $\alpha = 0.4$ (a practical value). This figure demonstrates that our receiver, using either $m = 8$ or $m = 16$, is able to achieve a near coherent performance whenever $N \geq 10$, i.e., whenever the burst duration exceeds 10 symbols. Meanwhile, the performance of a receiver using the tracking of [3] is close to coherent with $N = 100$, but, by $N = 10$, its performance has degraded to the point where reliable data detection is not attainable.

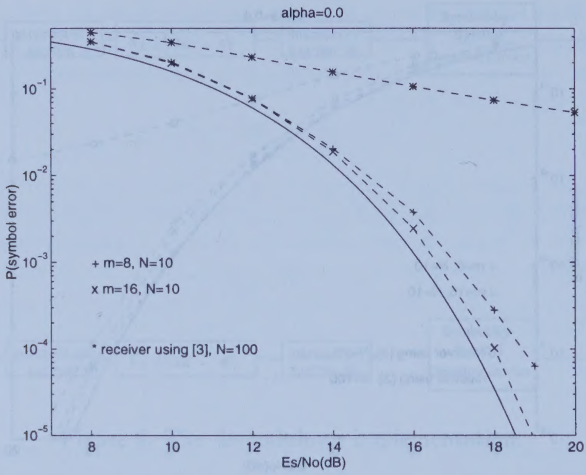


Figure 7: $P(\epsilon) - \frac{E_s}{N_o}$ for 8-PSK, $\alpha = 0.0$.

Figure 7 shows performances when $p_s(t)$ represents a *sinc* waveform, i.e., $\alpha = 0$. In this case, we observe that our receiver, using $m = 16$, is able to achieve a performance close to coherent for $N \geq 10$. A receiver employing the tracking of [3], on the other hand, is not able to achieve reliable data detection.

4.2 Complexity

The complexity of our receiver is measured in terms of computations per decoded symbol.

The receiver we introduce consists of two main computational components: m parallel branches and the CDU. The m parallel branches carry out, for each decoded symbol, m discrete time filterings, m symbol-by-symbol demodulations, and m likelihood computations (consisting of 2 multiplications and 1 addition each). Meanwhile, the CDU carries out m additions and $\frac{m-1}{N}$ of a comparison for each decoded symbol. Hence, the total complexity is in the order of m discrete time filterings and m symbol-by-symbol demodulations.

This complexity can be reduced whenever N, the burst length, exceeds 10. This reduced complexity is achieved by introducing a slight modification to our receiver implementation. We first describe the modified receiver implementation, and then present the corresponding complexity.

The reduced complexity novel receiver operates (conceptually) as follows. Over the first 10 symbols in each burst, the receiver operation corresponds to the implementation of Figure 5, using $N = 10$. Over the remaining symbols in the burst, the receiver simply applies the $j^{*\,th}$ matched filter and demodulator, where j^* corresponds to the index of the branch whose output was selected by the CDU over the first 10 symbols. This can be done because, when considering the first 10 symbols of the block, the simplified receiver has established which branch is best.

The computations carried out by the reduced-complexity novel receiver, in operations per decoded symbol, are: the m branches contribute $\frac{m\cdot 10+1\cdot(N-10)}{N}$ discrete time filterings and an equal number of demodulations and likelihood computations; the CDU introduces $\frac{m\cdot 10}{N}$ additions and $\frac{m-1}{N}$ comparisons. Hence, the total complexity is in the order of $\frac{m\cdot 10+1\cdot(N-10)}{N}$ discrete time filterings and demodulations per decoded symbol. The value of m is either 8 or 16, depending on the roll-off factor α. For example, with $\alpha = 0.4$ (in which case we can use $m = 8$) and $N = 100$, the total complexity is in the order of 1.7 discrete time filterings and demodulations per decoded symbol; with $N = 500$, this complexity drops to 1.14 discrete time filterings and demodulations per decoded symbol.

5 Conclusions

In this work, we have introduced a novel receiver for data detection in the presence of noise and timing offset. This receiver demonstrates many benefits when compared to receivers using a feedforward timing estimator, such as the widely used estimator of [3]. First, whenever the burst length is small, our receiver easily outperforms a receiver using the tracking of [3]. In fact, to the best of our knowledge, no receiver other than our own is able to achieve reliable data detection when the timing offset changes after every small burst length. Additionally, whenever the roll-off factor α is small, our receiver again easily outperforms a receiver using the estimator of [3]. With smaller roll-off factors becoming more and more common (to achieve higher bandwidth efficiency), this performance gain is an important one. Furthermore, our receiver is available at a low complexity. By using the reduced complexity implementation of our novel receiver, our receiver's complexity matches that of coherent over all but the first 10 symbols of each burst.

References

[1] E.A. Lee and D.G. Messerschmitt, *Digital Communication.* Boston, MA: Kluwer Academic Publishers, 1988.

[2] F.M. Gardner, 'A BPSK/QPSK timing error detector for sampled receivers', *IEEE Trans. Commun.*, Vol. 34, pp 423-429, May 1986.

[3] M. Oerder and H. Meyr, 'Digital filter and square timing recovery', *IEEE Trans. Commun.*, Vol. 36, pp 605-612, May 1988.

[4] C. W. Helstrom, *Statistical Theory of Signal Detection.* New York, NY: Pergamon Press, 1968.

[5] C.R. Nassar, *A novel receiver for data detection in the presence of nuisance parameters*, Ph. D. thesis, McGill University, Montreal, Quebec, 1996.

[6] C.R. Nassar and M.R. Soleymani, 'A quantized nuisance parameter space for data detection in the presence of rapidly changing nuisance parameters', to be submitted to *IEEE Trans. Commun.*

Signal Design for Improved Synchronization in Frequency Hopped Spread Spectrum CDMA Communications[*]

Lisa Ann Osadciw and Edward L. Titlebaum

University of Rochester
Rochester, NY 14627
(716)275-4061(office)/(716)275-2073(fax)
osadciw@ee.rochester.edu, tbaum@ee.rochester.edu

Abstract

This paper uses traceability information from the narrowband auto-ambiguity function to improve time and frequency synchronization of a frequency hopped spread spectrum signal in a CDMA communications system. Local oscillator mismatch and drift result in a frequency shift similar to a Doppler shift between spatially separate transmitter and receiver oscillators. Time and frequency shifts must be compensated for by acquisition and synchronization algorithms. The performance of these algorithms can be improved by taking advantage of the traceability of a FHSS signal. Traceability of a waveform is based on the number of grating lobes within 3 dB of the main peak in the narrowband auto-ambiguity function. The paper demonstrates that the linear congruence code family is the most traceable code family and the traceability of the waveform increases as the approximate *BT* product increases. Therefore, signal design can be used to increase traceabillity.

1. Introduction

Frequency hopped spread spectrum (FHSS) signals exhibit a traceability in the narrowband auto-ambiguity function[1,2] (NB AAF) which can be used to reduce the probability of false synchronization and acquisition time in communication systems. The usefulness of the NB AAF resides in the fact that the Doppler shift axis is analogous to the frequency mismatch between two spatially separated transmit and receive oscillators. Also the NB AAF quantifies the loss in detected signal power for time and frequency shifts in the transmitted and received signals. A pilot signal can be used to eliminate the Doppler shift caused by the movement of transmitter, receiver, or reflecting object within the channel but local oscillator mismatch and drift remain [3,4,5].

Synchronization and acquisition algorithms can compensate for this mismatch[6]. Current acquisition and synchronization techniques assume a "thumbtack" NB AAF. Therefore, grating lobes can lead to false alarm events in most acquisition schemes which are based on serial searches. First the simple serial search algorithm performs coarse synchronization by varying the detector's time and frequency offsets. This locates the mainlobe in the NB AAF or synchronizes the receiver to within a chip duration and approximate chip bandwidth of the transmitter. Fine synchronization is then performed with smaller variations in the time and frequency offsets. Synchronization is maintained by tracking the time and frequency mismatch and adjusting the offsets accordingly. Figure 1 illustrates the synchronization of the receiver and transmitter oscillators using the NB AAF of a waveform with a "thumbtack" NB AAF. The traceable FHSS

---

[*] *This work was sponsored in part by the BMDO/IST and managed by the Office of Naval Research under contract N00014-92-J-1933 and support by ASSERT grant N00019-95-I-1234.*

signals have NB AAFs that are the antithesis of a "thumbtack". The traceable patterns can then be used to reduce the acquisition and synchronization time in a FHSS CDMA system.

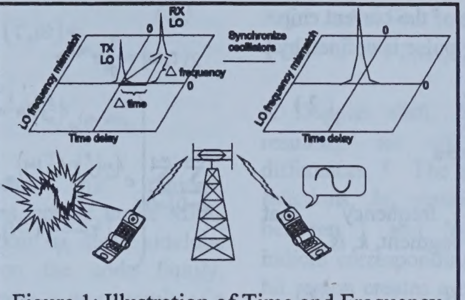


Figure 1: Illustration of Time and Frequency Synchronization of a Receiver

Local oscillator mismatch is a source of frequency mismatch which exists in every system and is specified by a system engineer in a frequency error budget. Typical budgets include initial calibration tolerance, stability over the operating temperature range, and long-term aging effects over 10 years[7]. These three factors are assumed to be independent of each other. Therefore, the total frequency tolerance is found by summing the three error budget tolerances. Error budgets can range from 5 parts per million (ppm) to 100 ppm. For a frequency error of only 5 ppm, the resulting oscillator mismatch can be as great as 20 kHz with a 2 Ghz carrier and approximate bandwidth of 1.25 Mhz. For an error tolerance of 100 ppm, the oscillator frequency error is 400 kHz. Higher error tolerances allow for less expensive oscillators and can reduce system cost.

This paper defines the traceability information found in the NB AAF. This signal design information can be used to improve signal acquisition and synchronization performance. The paper focuses on single-user issues by presenting only the auto-ambiguity function. Thus, multi-user interference, multipath, and jamming are ignored. The FHSS signal is assumed to be time-limited and approximately band-limited. M-ary Frequency Shift Keying, MFSK, is the modulation technique assumed.

The paper is divided into four sections. First, the theory section of the paper defines the traceability property of the NB AAF and identifies the information in the NB AAF needed to locate grating lobes for FHSS signals. Some examples are given in the third section. The final section contains conclusions and discusses future work.

2. Theory

In this section, the definition of the traceability property of the NB AAF is given based on the grating lobes originating from an exponential sum. First, the nomenclature defining the FHSS signals is introduced. Then the NB AAF is derived for the FHSS signals. An approach to finding the grating lobes in the NB AAF is described next. This description is followed by a discussion of the Costas condition, a key element in the NB AAF traceability property. This section concludes with a definition of traceability.

2.1 Frequency Hopped Signals

A FHSS signal varies its transmitted frequency periodically. For simplicity, all waveforms are uniformly weighted. The signal consists of *T* samples and N chips. The approximate total bandwidth is *B* radians [1]. The chip duration is *T/N* samples, and the chip bandwidth is *B/N* radians. . The hopping period is the reciprocal of chip duration.

The FHSS signal is expressed as

$$u(t) = \frac{1}{\sqrt{T}} \sum_{k=0}^{N-1} p\left(t - k\frac{T}{N} \right) e^{j[\omega_k t + \theta_k]} \quad (1)$$

where

ω_k = frequency of the current chip,

---

[1] *The time units can be changed from samples to seconds by dividing by the sampling rate in Hz.*

θ_k = initial phase of the current chip.

The rectangular gating pulse is defined by

$$p(t)=\begin{cases}1, & t\leq\dfrac{T}{N}\\ 0, & elsewhere\end{cases}\quad (2)$$

The instantaneous frequency that corresponds to the code segment, k, is

$$\omega_k=\omega_o+y_k\frac{B}{N}\,. \quad (3)$$

The definition of the placement operator in (3) depends on the code family. This paper discusses three code families: linear congruence codes, quadratic congruence codes, and hyperbolic congruence codes.

2.2 Narrowband Auto-Ambiguity Functions

The NB AAF can be defined in three ways: the matched filter response, the magnitude of the matched filter response, or the magnitude squared of the matched filter response. The matched filter response (MFR) is

$$\chi(\tau,\omega)=\int_{-\infty}^{\infty}u(t)u^*(t-\tau)e^{j\omega t}dt. \quad (4)$$

In this analysis, the NB AAF is defined as

$$\chi_{NB}(\tau,\omega)=\left|\chi(\tau,\omega)\right|. \quad (5)$$

(5) is sometimes referred to as the uncertainty function and is based on the real envelope of the MFR[2]. For clarity purposes, it should be stated that an analysis of the MFR instead of the NB AAF is used to locate the grating lobes. This is valid since the NB AAF is simply the magnitude of the MFR. Since the grating lobes are maximums, their locations remain the same in the MFR and NB AAF.

The MFR is rewritten using (1) for the signal gives

$$\chi(\tau,\omega)=\frac{1}{T}e^{j\omega_c\tau}\bullet \quad (6)$$

$$\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\left(\begin{array}{l} e^{j(\theta_k-\theta_l)}e^{j\left(\frac{BT}{N^2}\right)l(y_k-y_l)}\bullet \\ e^{j\left(y_k\left(\frac{B}{N}\right)\tau-l\left(\frac{T}{N}\right)(\omega)\right)}\bullet \\ \chi_p\left(\tau-\frac{T}{N}(k-l),\omega-\frac{B}{N}(y_k-y_l)\right)\end{array}\right)$$

where

$\chi_p(\tau,\omega)$ = the MFR of the rectangular pulse function,

τ = time shift in number of samples,

ω = frequency shift in radians per sample[2].

The MFR in (6) is a summation of N^2 terms. Each individual term can be referred to as subambiguity functions centered at the zeros of χ_p[8,9]. The area in the NB AAF where the subambiguity function resides is called the "hit" region[10]. The centers of the subambiguity functions are determined by

$$\tau=(k-l)\frac{T}{N} \quad (7)$$

and

$$\omega=(y_k-y_l)\frac{B}{N}\,. \quad (8)$$

Multiple subambiguity functions can share the same center location in the NB AAF. This is referred to as a multiple hit region. For example at the origin, N subambiguity functions, corresponding to $k = l$, coincide forming the mainlobe region. The remaining subambiguity functions, not centered at the origin, have minimal effect on the grating lobes in this mainlobe region. For the mainlobe region, (6) can be approximated using N terms by

---

[2] *This assumes the congruence code is for the finite field. For the group, the summation in (6) begins at 1 reducing the total number of subambiguity functions to (N-1)^2.*

$$\widetilde{\chi}(\tau,\omega)=\frac{1}{T}\chi_p(\tau,\omega)\bullet \sum_{k=0}^{N-1} e^{j\left(\left(\frac{B}{N}\right)y_k\tau-\left(\frac{T}{N}\right)k\omega\right)}. \tag{9}$$

Less significant "hit" regions, referred to as sidelobes, appear in the MFR. Multiple hits can occur in these sidelobe regions depending on the code family. Linear congruential codes is an example of a code family with significant sidelobe regions. In order to analyze the other hit regions, it is useful to introduce a change of variables or

$$\Delta=(k-l) \tag{10}$$

and

$$y_k-y_l=y_{k-\Delta}-y_k \quad . \tag{11}$$

In (10), the Δ variable can take on values ranging from 0 to N-1 which ignores the negative values of τ.

An approximation of the MFR composed of the subambiguity functions centered Δ is

$$\widetilde{\chi}(\tau,\omega)=\frac{1}{T}e^{j\omega_c\tau}\bullet \sum_{k=0}^{N-1-\Delta}\left(\begin{array}{l} e^{j\left(y_{k-\Delta}\frac{B}{N}\tau-k\frac{T}{N}\omega\right)}\bullet \\ \chi_p\left(\tau-\frac{T}{N}\Delta,\omega-\frac{B}{N}(y_{k-\Delta}-y_k)\right)\end{array}\right). \tag{12}$$

The subambiguity functions centered about a consistent time shift in (12) are distributed in Doppler by (11). The key to finding an approximation of the MFR for a single hit region in τ and ω is to isolate the terms in (12) which are centered about a selected frequency shift. The set of k values for a chosen Δ time shift is

$$k\in S=\{0,1,\ldots,N-1-\Delta\}\cdot \tag{13}$$

The corresponding placement differences are

$$y_{k-\Delta}-y_k\in Y\subset \{-N+1,-N+2,\ldots,0,\ldots,N-2,N-1\}. \tag{14}$$

A Doppler shift, δ, is chosen from the resulting set of placement operator differences, Y. The number of subambiguity functions, M, comprising a hit region is between 1 and N-1-Δ. Separating the indices corresponding to the k values in the hit region creates a new set of index values which is a subset of (13) is created giving

$$k_g\in S_g\subset S\cdot \tag{15}$$

The MFR for any hit region is approximated by

$$\widetilde{\chi}(\tau,\omega)=\frac{1}{T}e^{j\omega_c\tau}\bullet \sum_{g=1}^{M}\left(e^{j\left(y_{k_g-\Delta}\left(\frac{B}{N}\tau\right)-k_g\left(\frac{T}{N}\omega\right)\right)}\bullet\chi_p\left(\tau-\frac{T}{N}\Delta,\omega-\frac{B}{N}\delta\right)\right). \tag{16}$$

For simplification purposes, (16) can be rewritten as

$$\widetilde{\chi}(\tau,\omega)=\frac{1}{T}e^{j\omega_c\tau}\bullet e^{-j\Delta\left(\frac{T}{N}\right)\omega}\bullet \sum_{g=1}^{M}\left(\begin{array}{l} e^{j\left(y_{k_g}\left(\frac{B}{N}\tau\right)-k_g\left(\frac{T}{N}\omega\right)\right)}\bullet \\ \chi_p\left(\tau-\frac{T}{N}\Delta,\omega-\frac{B}{N}\delta\right)\end{array}\right). \tag{17}$$

Thus, the approximation of the NB AAF for a "hit" region specified by Δ and δ is

$$\widetilde{\chi}_{NB}(\tau,\omega)=\left|\frac{1}{T}e^{j\omega_c\tau}\right|\bullet \left|\chi_p\left(\tau-\frac{T}{N}\Delta,\omega-\frac{B}{N}\delta\right)\right|\bullet \left|\sum_{g=1}^{M}e^{j\left(y_{k_g}\left(\frac{B}{N}\tau\right)-k_g\left(\frac{T}{N}\omega\right)\right)}\right|. \tag{18}$$

χ_p controls the amplitude of the grating lobes while the exponential summation is the source of the grating lobes[11].

2.3 Approach to Finding Grating Lobes Within Hit Region

The location of the grating lobes within any of the hit regions is determined by analyzing the exponential sum or

$$\chi_e(\tau,\omega) = \sum_{g=1}^{M} e^{j\left(y_{k_g}\left(\frac{B}{N}\right)\tau - k_g\left(\frac{T}{N}\right)\omega\right)} . \qquad (19)$$

A grating lobe is formed where the phases in (19) are integer multiples of 2π for all values of g. There must be a minimum of 3 subambiguity functions coinciding within the hit region for grating lobes to be found this way. If the hit consists of a single subambiguity function, the hit region matches the NB AAF of the single pulse. If two subambiguity functions coincide, ridges are formed at the beat frequencies.

Defining the phase function in the exponential of (19) as

$$\Phi\left(k_g,\tau,\omega,\Delta;y_{k_g}\right) = \left(y_{k_g}B\tau - k_gT\omega\right)\left(\frac{1}{N}\right). \qquad (20)$$

Lines of constant phase are derived from (20) . The grating lobe peaks occur where the M lines of constant phase intersect. Setting (20) equal to any multiple of 2π gives

$$\left(y_{k_g}B\tau - k_gT\omega\right)\left(\frac{1}{N}\right) = m_g 2\pi \ . \qquad (21)$$

where

m_g = any positive or negative integer.

Rewriting this as the equation of a line results in

$$\omega = \left(\frac{y_{k_g}}{k_g}\right)\left(\frac{B}{T}\right)\tau + \frac{N2\pi m_g}{k_gT} \ . \qquad (22)$$

(22) describes a line of constant phase for each $\{k_g, y_{\Delta+g}\}$ pair in (τ,ω) space within the hit region of interest. The intersection of these lines locates a grating lobe. The M equations from (22) can be used to form a linear matrix of equations. For each set of m_g parameter values, the solution of this matrix specifies a grating lobe location. The parameters must be chosen so that the solution remains within the hit region. Recall that the NB AAF of the chip determines the area which the hit region occupies.

2.4 Costas Condition

One observation to be made from the summation of the exponential terms in (18) is that as the time-bandwidth product increases, more grating lobes appear. This observation is related to the Costas condition [12].

The Costas condition is derived by analyzing the MFR along the τ axis. The first grating lobe resulting from the exponential sum in (19) is located on the τ axis at

$$\tau = \frac{2\pi N}{B}. \qquad (23)$$

The grating lobe peak is eliminated if

$$BT \geq N^2 2\pi . \qquad (24)$$

As BT increases, more grating lobes appear along the τ axis. The Costas condition is met if the waveform parameters of a FHSS signal are chosen so that

$$BT = N^2 2\pi . \qquad (25)$$

This condition applies to full and non-full frequency hop codes.

Similarly, the first grating lobe can be found along the Doppler axis by setting ω to 0 in (16). The first grating lobe peak occurs at

$$\omega = \frac{2\pi N}{T} \ . \qquad (26)$$

The NB AAF of the rectangular pulse is not finite along the ω axis like the NB AAF is along the τ axis. Therefore, there is no guarantee that grating lobes are eliminated in Doppler if the Costas condition is met. As the examples demonstrate, some code families have numerous grating lobes in the NB AAF when the Costas condition is met.

2.5 Traceability Property of Narrowband Auto-Ambiguity Functions

The traceability property as is the Doppler tolerance property of a waveform's NB AAF is the antithesis of a thumbtack NB AAF. Doppler tolerance refers to signals that support the detection of a signal in the presence of Doppler shift. For example, a chirp signal is Doppler tolerant. The traceability property allows for faster acquisition and synchronization of a signal by forming predictable and detectable grating lobes in the NB AAF. A system can trace a pattern to the primary peak of the mainlobe using knowledge of the grating lobe structure. One quantity that is used to measure traceability in this paper is the number of predictable grating lobes within 3 dB of the main peak in a hit region.

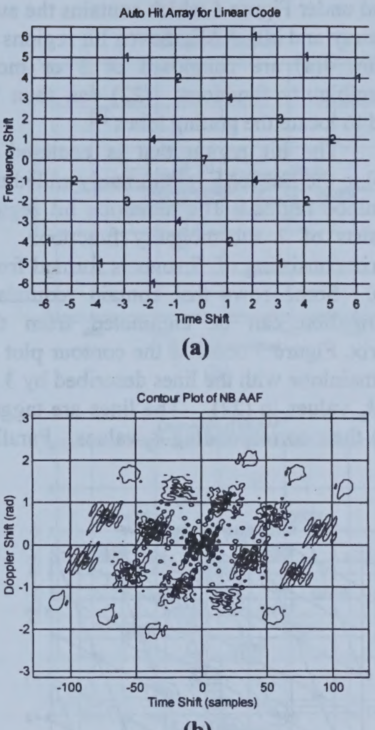
The traceability of the waveform increases as the *BT* product increases. If the Costas condition is met, there are no grating lobes along the τ axis. However, the Costas condition does not lead to any assumptions concerning grating lobes off the τ axis. For many code families, no grating lobes appear for the Costas condition. Traceability is measured and compared in the next section for a number of FHSS signal examples.

3. Examples of Various Frequency Hopped Signals

The traceability of the linear congruence code family is compared with the traceability of the hyperbolic congruence code family and quadratic congruence code family in this section.

3.1 Linear Congruential Codes

Two beneficial features of the family of linear congruence (LC) code are its ideal cross-ambiguity function which minimizes multi-user interference and the ease of code generation. For a LC frequency hop code, the placement operator, y_k, is defined as



(a)

(b)

Figure 4: (a) Auto Hit Array for LC Code (p=7,a=3,B=2.4050,T=128) (b) NB AAF

$$y_k \equiv ak \pmod{p} \qquad (33)$$

where

p = the prime number for the finite field, $GF(p) = \{0, 1, ..., p-1\}$,
a = the multiplier, $a \in GF(p) - \{0\}$,
k = index $k \in GF(p)$ [3] [9].

A tractable example with a small code length demonstrates the effectiveness of using (22) to locate the grating lobes. The first example meets the Costas condition in (30). The parameter values are listed under Figure 4 which contains the auto hit array and NB AAF. Seven hit regions in Figure 4(a) are composed of 3 or more subambiguity functions. (22) can then be used to locate the grating lobes [8].

---

[3] *The placement operator is based on modular arithmetic over a finite field of some prime number p, GF{p}. The "≡" symbol is read "is congruent to"* [17].

The hit region that is centered at (0,0) in the NB AAF is commonly called the mainlobe region. The mainlobe hit region consists of 7 subambiguity functions. A matrix consisting of 7 rows is formed from (22). Some rows that contain redundant information can be eliminated from the matrix. Figure 5 contains the contour plot of the mainlobe with the lines described by 3 of the k_g values in (22). The lines are tagged with their corresponding k_g values. Parallel lines are formed as the m_g parameter is varied. There are 8 grating lobes within 3 dB of the primary peak at (0,0) in Figure 5. Thus, this LC code is traceable even when the Costas condition is met.

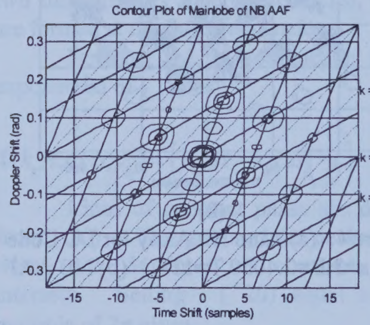


Figure 5: Mainlobe of NB AAF with Lines Predicting Grating Lobe Locations

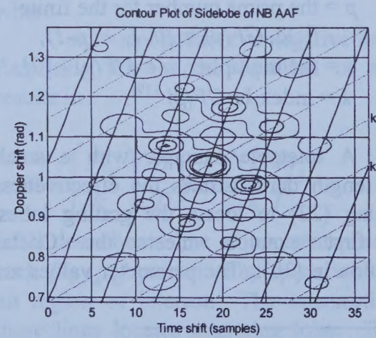


Figure 6: Sidelobe of NB AAF with Lines Predicting Grating Lobe Locations

There are 6 sidelobe regions containing more than three hits which can be analyzed using (22). From Figure 4 (a), a sidelobe with 4 hits occurs for an index shift or Δ time shift of 1 and δ Doppler shift of 3. The center of the sidelobe region maps to (17, 1.02) in the (τ, ω) plane. Figure 6 illustrates the grating lobes including the intersecting line pattern from (22). The sidelobes in the NB AAF are also traceable for the LC code.

3.2 Hyperbolic Congruential Codes

In this section, two examples are provided for the hyperbolic congruence (HC) code family. These examples demonstrate how the traceability property of the NB AAF is a function of the *BT* product as well as code family. The first example meets the Costas condition, and the second example exceeds the Costas condition by increasing the bandwidth. The HC code family is a full code which has lower sidelobe levels in the auto-ambiguity function than the LC code family. Thus, the HC code approaches a "thumbtack" by nearly spreading the hits uniformly in the auto hit array. In Figure 7, the auto hit array for an HC code illustrates that the maximum number of subambiguity functions coinciding in a sidelobe hit region is 2. The hyperbolic congruential code is defined by

$$y_k \equiv \frac{a}{k} (\bmod p) \qquad (34)$$

where

p = the prime number of the finite field, $GF(p) = \{0, 1, \ldots, p-1\}$,
a = the multiplier, $a \in GF(p) - \{0\}$,
k = index $k \in GF(p) - \{0\}$[9].

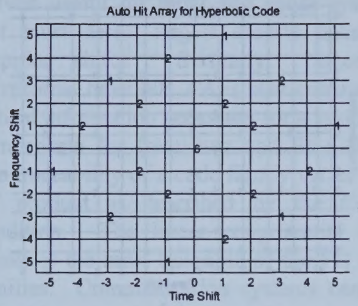


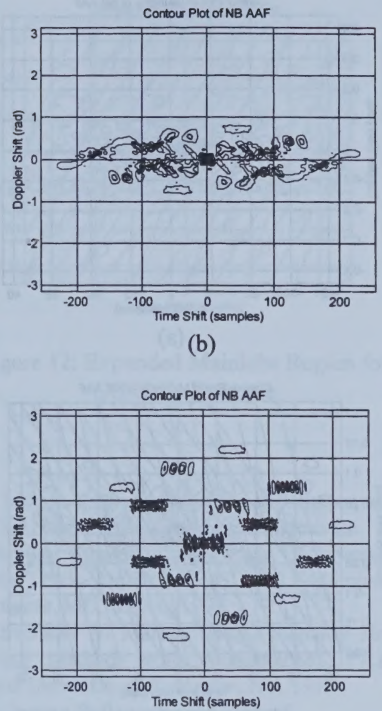
Figure 7: Auto Hit Array of HC Code

The auto hit array and NB AAFs for the HC code example are illustrated Figure 8. The auto hit array which is independent of the *BT* product is given in Figure 8(a). Figure 8(b) is a contour plot of the NB AAF for the HC code that meets the Costas condition. The actual parameter values are listed below the figures. Figure 8(c) contains the NB AAF of a waveform with a higher approximate bandwidth.

This example demonstrates that the HC code family is slightly traceable. As shown in Figure 9, two grating lobes appear in the mainlobe for the first example which meets the Costas condition. Six grating lobes are formed that exceed the 3 dB limit in the second example which has a higher *BT* product. Two of the grating lobes appear along the time shift axis at +/- 15 as shown in Figure 10. Figure 10 (a) is a contour plot of the entire mainlobe region, and Figure 10 (b) is an expanded view of the mainlobe region including the lines predicting the grating lobe locations. Thus, increasing the *BT* product increases traceability.

3.3 Quadratic Congruential Codes

The quadratic congruence (QC) code family has the ability to reduce interference from colored noise or clutter. The auto-ambiguity function achieves this by minimizing the signal energy along a strip parallel to the zero Doppler axis[14]. The QC code is not a full code so that some frequency channels are repeated and others skipped. This leads to hit regions along the



(b)

(c)

Figure 8: NB AAF of Hyperbolic Congruence Code (b) NB AAF (p=7, a=3, B=0.8836, T=256) (c) NB AAF (p=7, a=3, B=2.6508, T=256)

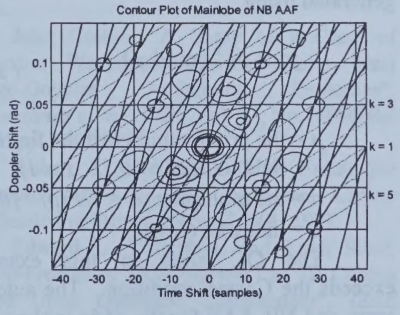
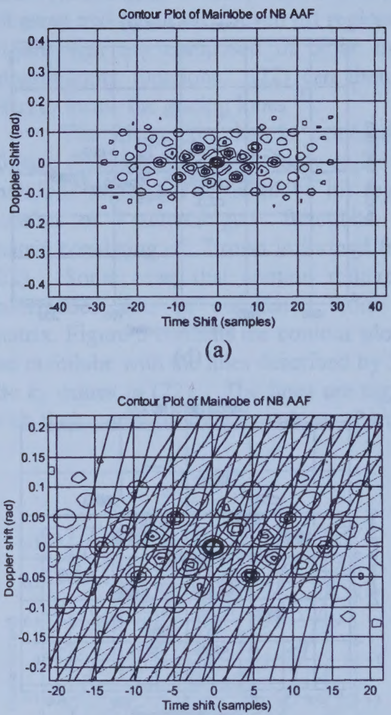


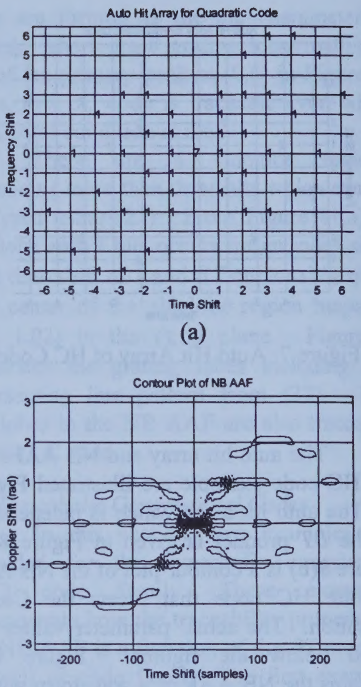
Figure 9: Mainlobe of NB AAF for HC Code Meeting Costas Condition



(a)

(b)

Figure 10: Mainlobe of NB AAF for HC Code Exceeding Costas Condition
(a) entire mainlobe (b) expanded region



(a)

(b)

Figure 11: Quadratic Congruence Code Example (p=7, a=3, B=2.4053, T=256)

(a) Auto Hit Array (b) NB AAF

time shift axis. The placement operator is generated using

$$y_k \equiv ak^2 \pmod p \qquad (35)$$

where

p = the prime number of the finite field, $GF(p) = \{0,1, \ldots, p-1\}$,
a = the multiplier, $a \in GF(p) - \{0\}$,
k = index $k \in GF(p)$[9,14].

The QC code in this example exceeds the Costas condition. The auto hit array and NB AAF for the code is shown in Figure 11. The hit regions along the time shift axis are clearly visible. In Figure 12, the NB AAF of the mainlobe region is shown. The region is expanded for readability, and the lines that predict the grating lobe positions are superimposed. There are 2 grating lobes in this example exceeding 3 dB. The QC code family is only traceable for higher *BT* products.

4. Conclusions and Future Work

The "thumbtack" auto ambiguity function is optimum for communication, sonar, and radar systems that have apriori knowledge of the transmitted waveform and channel distortions. Expensive signal processors which measure and compensate for frequency shifts and time delays can compensate for any lack of information. However, traceability is a signal design tool which can be used when information is missing. Doppler tolerant waveforms have proven useful for radar and sonar systems that must detect targets despite unknown Doppler shifts. Similarly, traceable waveforms have NB AAFs that contain the critical information necessary for a system to compensate for frequency and time shifts. The traceability of a code family varies with *BT* product as described by the Costas condition. The linear congruential code family is the most traceable of the FH code families. Communication systems can use traceable signals to speed up signal acquisition and synchronization in the receiver[15].

In future work, the traceability of signals in nonideal conditions must be addressed. Multi-user interference, frequency fading, and jamming are issues that need to be studied. Acquisition and synchronization schemes must also be developed to take full advantage of the traceability property of some FHSS waveforms.

References

[1] P. M. Woodward, *Probability and Information Theory, with Applications to Radar*, Pergamon Press, Inc., NY, 1953.

[2] August Rihaczek, *Principles of High-Resolution Radar*, McGraw-Hill Book Company, 1969.

[3] Mark Hahm, and Edward Titlebaum, "Wideband Signal Design Issues and Local Oscillator Drift", *Proc. 1996 Princeton CISS Conference*, Princeton, NJ, April 1996, pp. 1151-1152.

[4] Theodore S. Rappaport, *Wireless Communications: Principles & Practice*, Prentice Hall PTR, Upper Saddle River, NJ, 1996.

[5] Vijay K. Garg, Kenneth Smolik, and Joseph E. Wilkes, *Applications of CDMA in Wireless/Personal Communications,* Prentice Hall PTR, Upper Saddle River, NJ, 1997.

[6] Marvin K. Simon, Jim K. Omura, Robert A. Scholtz, and Barry K. Levitt, *Spread Spectrum Communications Handbook,* McGraw-Hill, Inc. 1994.

[7] Robert R. Zeigler, *Specifying Crystals and Oscillators*, Cardinal Components, Inc., http://cnct.com/home/cardinal, 1997.

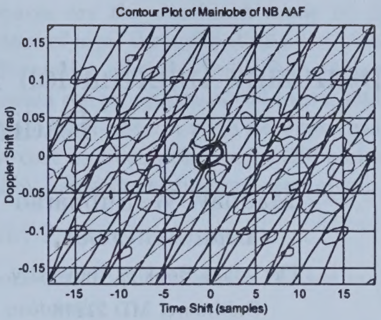


Figure 12: Expanded Mainlobe Region for QC Code Example

[8] Lisa A. Osadciw, Edward L. Titlebaum, and Mark Hahm, "Mainlobe Characteristics of Ambiguity Functions for Linear Congruential Codes", *Proc. Baltimore 1997 CISS Conference,* Baltimore, MD, March, 1997.

[9] Svetislav V. Maric', "New Frequency Hop Coding Methods with Applications", *Ph.D. Thesis University of Rochester*, NY, 1990.

[10] Jerome Bellegarda and Edward Titlebaum,"The Hit Array: An Analysis Formalism for Multiple Access Frequency Hop Coding", IEEE Transactions on Aerospace and Electronic Systems, Vol. 27, No. 1, January 1991, pp. 30-39.

[11] Bernard Steinberg, *Principles of Aperture and Array System Design,* John Wiley & Sones, 1976.

[12] John Costas, "A Study of a Class of Detection Waveforms Having Nearly Ideal Range-Doppler Ambiguity Properties", *Proceedings of the IEEE*, Vol. 72, No. 8, August 1984, pp. 996-1009.

[13] Robert J. McEliece, *Finite Fields for Computer Scientists and Engineers*, Kluwer Academic Publishers, Norwell, MA, 1987.

[14] Edward L. Titlebaum, and Leon Sibul, "Time-Frequency HOP Signals Part II: Coding Based Upon Quadratic Congruences", *IEEE Transaction on Aerospace and Electronic Systems,* vol-AES-17, No. 4, July 1981, pp. 494-499.

[15] Mark David Hahm, "Receiver Issues Related to Spread Spectrum Communications", *Ph.D. Thesis University of Rochester*, NY, 1997.

Optimizing Information Efficiency in a Mobile Packet Radio Network

Madhavi W. Subbarao
Department of ECE
The Johns Hopkins University
Baltimore, MD 21218

Brian L. Hughes
Department of ECE
North Carolina State University
Raleigh, NC 27695

Abstract — We consider a direct-sequence multihop packet radio network using the slotted ALOHA channel access protocol in a Rayleigh fading environment. For such networks, we are interested in optimally choosing the transmission range, code rate, and slotted ALOHA transmission probability to be used by each node. Performance measures of local throughput and *information efficiency* are used to analyze the network. Information efficiency is a measure of "expected forward progress" that captures the *amount of information* being forwarded *per unit bandwidth.* Maximizing information efficiency yields the optimum desired network parameters. We show how the information efficiency of the low-cost mobile packet radio network can be improved approximately 14-fold by using the optimum parameters.

I. INTRODUCTION

In a mobile packet radio network, it is not always advantageous for the source to transmit directly to the destination, even though it is within transmission range. Instead, it might prove beneficial for the packet to take a series of short hops, thereby achieving spatial reuse and increasing network throughput [1]. Two important issues for efficient network performance are to find the optimum transmission range and optimum code rate to be used by each node in the network. The optimum combination is a tradeoff between the distance covered by a transmission and the amount of data transmitted.

For narrow-band packet radio networks with zero capture, it was shown that the optimum transmission range is such that, on average the number of nodes within range is between six and eight [2, 3, 4, 5, 6]. This problem was also analyzed for a direct-sequence (DS) spread-spectrum packet radio network [7, 8]. In [9], we addressed this problem *coupled* with finding the optimum code rate for an asynchronous frequency-hopping multihop packet radio network.

The optimum transmission ranges in [3, 5, 6, 7, 8] were found by maximizing the *expected forward progress.* Given that a transmission is successful, the expected forward progress is the distance between the transmitter and receiver projected onto a line drawn between the transmitter and final destination. This metric is a measure of *distance* and does not account for the amount of data transmitted. In [9], we introduced a new performance measure, *information efficiency,* which is a measure of expected forward progress that captures the *amount of information* being forwarded *per unit bandwidth.*

In this paper, we find the optimum transmission range *coupled* with finding the optimum code rate for a DS multihop packet radio network using the slotted ALOHA channel access protocol. We describe the underlying system model, and present the packet error probability of the system under the assumptions that the desired signal undergoes Rayleigh fading and that ground waves are the dominant mode of signal propagation. We describe the performance measure of local throughput and our measure of *information efficiency.* By maximizing the information efficiency subject to a tolerable local throughput, we find the optimum transmission range, optimum code rate, and optimum slotted ALOHA transmission probability for each node in a DS network. Finally, we analyze the information efficiency of the low-cost mobile packet radio [10] and show how the information efficiency can be improved by using the optimum parameters.

II. SYSTEM MODEL AND PACKET ERROR PROBABILITY

We consider a DS multihop packet radio network with Rayleigh fading, Poisson random traffic, and the slotted ALOHA channel access protocol.

Consider a transmitter communicating via DS modulation with a receiver at a distance of r_0. The input data stream is divided into an *in-phase* stream and a *quadrature* stream, which are modulated with separate in-phase and quadrature spreading sequences respectively. After spreading, the in-phase stream is multiplied by a cosine function of a carrier wave, and the quadrature-stream is multiplied by the corresponding sine function. The two signals are then combined and sent on the radio frequency carrier. As the transmitted signal propagates to the receiver, its power decays with distance as $K^2 r_0^{-\eta}$ (equivalently, its amplitude decays as $K r_0^{-\eta/2}$), where K is a constant and η is the *path loss exponent* (typically $2 \leq \eta \leq 4$). The signal amplitude is also subject to random fluctuations caused by Rayleigh fading.

At the receiver, the desired signal is corrupted by interference from other active nodes in the network. We model the locations of these nodes as distributed randomly over the plane according to a two-dimensional Poisson point process with parameter λ, i.e.,

$$\mathrm{P}[k \text{ nodes in area } A] = \frac{(\lambda A)^k \exp(-\lambda A)}{k!}, \quad k = 0, 1, 2, \ldots,$$

where λ is the average number of nodes per unit area. The network operates under a heavy traffic condition (i.e., nodes always have packets waiting to be transmitted) and channel access is governed by the slotted ALOHA protocol, so each node transmits a packet with probability p in each slot. We assume that interfering nodes use the same modulation scheme as the transmitter and all nodes transmit with power P_T. Given the locations of the interfering users, the total interfering signal is assumed to be a Gaussian random process [7].

Let P_0 be the received power due to the transmitter, P_I be the power of the multiple-access interference, and B be the power of the background noise. The probability of successful packet reception $P_P(C)$ is then given by

$$P_P(C) = \Pr\left[\frac{P_0}{P_I + B} > b\right], \tag{1}$$

where b is the signal-to-noise ratio (SNR) threshold required for the desired probability of error. Threshold b depends on the particular signaling scheme and error-correction coding scheme used. In [8], it was shown that $P_P(C)$ for a DS network in a Rayleigh fading environment is given by

$$P_P(C) = e^{-\mu-\beta H}, \tag{2}$$

where

$$\begin{aligned}\mu &= \frac{b}{SNR_0},\\ \beta &= pN\,(b/L)^{2/\eta},\\ H &= \frac{2\pi}{\eta}\csc[2\pi/\eta]e^{2(1/\eta)^2},\end{aligned}$$

where SNR_0 is the SNR due to background noise only, p is the slotted ALOHA nodal transmission probability, L is the number of chips/symbol, and

$$N = \lambda\pi r_0^2 \tag{3}$$

is the average number of terminals closer to the receiver than the transmitter.

III. PERFORMANCE MEASURES

Measures of *local throughput* and *information efficiency* are used to analyze the performance of the network. Local throughput ξ is defined as the rate at which a node successfully transmits packets. In [7], it was shown that ξ, measured in packets per slot, is given by

$$\xi = (1-p)(1-e^{-p})P_P(C), \tag{4}$$

where $P_P(C)$ is the probability of correct packet reception.

In [9], we introduced information efficiency as a measure of "expected forward progress" that accounts for the *amount of information* being forwarded *per unit bandwidth*. Information efficiency for a network is given by

$$\zeta = r_0\,\xi\,\frac{R}{TW}, \tag{5}$$

where R is the code rate in bits/channel use, T is the symbol duration, W is the system bandwidth, and r_0 is the distance covered by a transmission, i.e., distance between transmitter and intended receiver. The quantity ζ is a communications analog of *momentum* in classical physics and yields the number of bit-meters per second forwarded per unit bandwidth. It may be interpreted as the average velocity at which a bit travels in the network per unit bandwidth. Similar to other definitions of expected forward progress, information efficiency *increases* as the probability of correct packet reception increases, and *decreases* with an increase in the number of hops a packet must take in order to reach its destination. (More hops implicitly leads to greater traffic congestion, and thus a decrease in the probability of correct packet reception.) Using (3) and substituting for r_0 in (5), and multiplying by $\sqrt{\lambda}$ to remove the dependence on the density of terminals per unit area, we obtain

$$\sqrt{\lambda}\zeta = \sqrt{\frac{N}{\pi}}\frac{R}{L}\,\xi\;, \tag{6}$$

which is the normalized information efficiency for a DS system with bandwidth $W = L/T$.

IV. OPTIMIZING DESIGN PARAMETERS

The design parameters of interest include transmission range measured in terms of N, code rate R, slotted ALOHA transmission probability p, and the SNR threshold b. Maximizing the information efficiency will determine the optimum parameters.

To perform this multivariate optimization, we first find the partial derivative of $\sqrt{\lambda}\zeta$ with respect to N and set it to zero, obtaining

$$\frac{\partial(\sqrt{\lambda}\zeta)}{\partial N} = \frac{R}{L\sqrt{N\pi}}(1-p)(1-e^{-p})\, e^{-\frac{b}{SNR_0}-C_0 N}\,[1/2 - C_0 N] = 0, \tag{7}$$

where $C_0 = p(b/L)^{2/\eta}H$. Solving (7) for N yields

$$N_{opt} = \frac{1}{2C_0}, \tag{8}$$

which is the optimum number of nodes that should be in transmission range. Substituting N_{opt} in (6), we obtain

$$\sqrt{\lambda}\zeta = \frac{R\, e^{-\frac{b}{SNR_0}-\frac{1}{2}}}{L\sqrt{2\pi(b/L)^{2/\eta}H}}\;\frac{(1-p)(1-e^{-p})}{\sqrt{p}}, \tag{9}$$

which is concave on $[0,1]$ in p and achieves its maximum at $p \approx 0.271$. Hence, the optimal transmission probability is

$$p_{opt} = 0.27 \tag{10}$$

regardless of the code rate and transmission range.

Since the total noise given the locations of the interfering users, i.e., multiple-access interference plus background noise, is a Gaussian random process, the DS network with a fixed processing gain is essentially a discrete-time additive Gaussian noise channel. Thus, Shannon's limit [11] from classical communications provides the constraint

$$\gamma \geq \frac{2^R - 1}{R}, \tag{11}$$

where R is the code rate in bits/channel use, and $\gamma = bL/R$ is the required E_b/N_0 threshold. (See the Appendix for details.)

Using N_{opt} and p_{opt}, and rewriting $\sqrt{\lambda}\zeta$ in terms of γ, we obtain

$$\sqrt{\lambda}\zeta = \frac{C_1\, R\, e^{-\frac{\gamma R}{L\, SNR_0}-\frac{1}{2}}}{L\sqrt{2\pi H}}\left(\frac{L^2}{\gamma R}\right)^{1/\eta}, \tag{12}$$

where $C_1 = (1-p_{opt})(1-e^{-p_{opt}})/\sqrt{p_{opt}} = 0.33$. Clearly (12) increases as γ decreases. However, γ is constrained by (11), and thus, the maximum value of $\sqrt{\lambda}\zeta$ is achieved at $\gamma_{opt} = (2^R-1)/R$. Substituting γ_{opt} in (12) yields

$$\sqrt{\lambda}\zeta = \frac{C_2\, R\, \exp\left[-\frac{(2^R-1)}{L\, SNR_0}\right]}{(2^R-1)^{1/\eta}}, \tag{13}$$

where

$$C_2 = \frac{C_1\, L^{2/\eta-1}}{\sqrt{2\pi e H}}. \tag{14}$$

The unique value of R that maximizes (13) R_{opt} can be easily found using numerical methods for optimization. The optimum SNR threshold is then

$$b_{opt} = \frac{2^{R_{opt}} - 1}{L}. \tag{15}$$

V. NETWORK ANALYSIS

An important consideration in network performance is to require an *acceptable level of local throughput* ξ_0. To this end, we analyze the information efficiency subject to a local throughput of at least 90% of the maximum attainable. Consider the low-cost mobile packet radio network described in [10]. The system parameters of this network are the following: $L = 128$, $b = 6.6$ dB, $R = 1/2$, required noise level $\mu = -20$ dB, $\eta = 4$, and $\beta/p = 1.18$. The constrained information efficiency $\sqrt{\lambda}\zeta_{\{\xi>\xi_0\}}$ of the low-cost mobile packet radio network is $\sqrt{\lambda}\zeta_{lc|\{\xi>\xi_0\}} = 3.4 \times 10^{-4}$ bit-m/s per Hz with a transmission range encompassing $N \approx 1$ node.

Using the analysis from Section IV and numerical methods for optimization to determine the value of R which maximizes (13), we find that the optimum code rate for the low-cost mobile packet radio network is $R = 5.6$ bits/channel use with corresponding SNR threshold $b = -4.3$ dB. Thus, the maximum attainable constrained information efficiency of the low-cost packet radio network is $\sqrt{\lambda}\zeta_{lc-opt|\{\xi>\xi_0\}} = 7.7 \times 10^{-3}$ bit-m/s per Hz with a transmission range encompassing $N \approx 4$ nodes. A practical coding scheme which comes close to realizing this theoretic limit is 16QAM trellis-coded modulation with 16 states, code rate $R = 3.0$ bits/channel use, and SNR threshold $b = -7.1$ dB [12, 13]. Using this coding scheme, the constrained information efficiency of the low-cost mobile packet radio network can be improved ≈ 14 times, i.e., $\sqrt{\lambda}\zeta_{lc-improved|\{\xi>\xi_0\}} = 4.6 \times 10^{-3}$ bit-m/s per Hz with a transmission range encompassing $N \approx 4$ nodes.

In Figures 1, 2, and 3 we plot the constrained information efficiency $\sqrt{\lambda}\zeta_{\{\xi>\xi_0\}}$ vs. code rate R and transmission range (measured in terms of number of nodes N) for the DS system with different processing gains L. For (R, N) pairs that do not achieve at least 90% local throughput, the information efficiency is considered to be zero. As L decreases, $\sqrt{\lambda}\zeta_{\{\xi>\xi_0\}}$ increases, however, the number of valid (R, N) pairs decreases, ultimately resulting in smaller transmission ranges. A striking observation is that the optimum code rate is robust against varying processing gains L and remains around $R = 5$ bits/channel use.

VI. APPENDIX

A. SNR Threshold

Let γ be the $E_b/N_{0_{eff}}$ threshold required for correct packet reception, where $N_{0_{eff}}/2$ is the power spectral density of the effective background noise. Then

$$\begin{aligned} E_b/N_{0_{eff}} &= \frac{P_0}{N_{0_{eff}} D} \\ &= \frac{P_0}{(N_0 + P_I/W)D} \\ &= \frac{P_0 W}{(W N_0 + P_I)D} \end{aligned}$$

$$\frac{P_0 W}{(W N_0 + P_I)D} \geq \gamma$$

$$\frac{P_0}{WN_0 + P_I} \geq \frac{\gamma D}{W}, \tag{16}$$

where D is the data rate in bits/sec, P_0 is the received signal power, P_I is the power of the multiple-access interference, W is the system bandwidth, and $N_0/2$ is the power spectral density of the thermal noise.

Since the left hand side of (16) is the SNR, the right hand side of (16) is the SNR threshold b, i.e., $b = (\gamma D)/W$. For the DS signaling scheme described in Section II, $D = R/T$ and $W = L/T$, where L is the number of chips per symbol. Then

$$b = \frac{\gamma R}{L}. \tag{17}$$

B. Information-Theoretic Bounds for DS Signaling

For the transmission scheme described in Section II, the total power is split between the in-phase and quadrature orthogonal data streams of power $P' = P/2$, each of which is bounded by Shannon's limit for a discrete-time additive white Gaussian noise channel [11]. The combined data rate D of both streams is then

$$\begin{aligned} D &< 2 \cdot \frac{1}{2T} \log_2 \left(1 + \frac{2P'T}{N_0}\right) = \frac{1}{T} \log_2 \left(1 + \frac{PT}{N_0}\right) \\ \longrightarrow R &< \log_2 \left(1 + \frac{E_b}{N_0} \cdot R\right) \\ \longrightarrow \frac{E_b}{N_0} &> \frac{2^R - 1}{R}, \end{aligned} \tag{18}$$

where $R = DT$ is the code rate in bits/channel use, and $E_b/N_0 = P/(N_0 D)$.

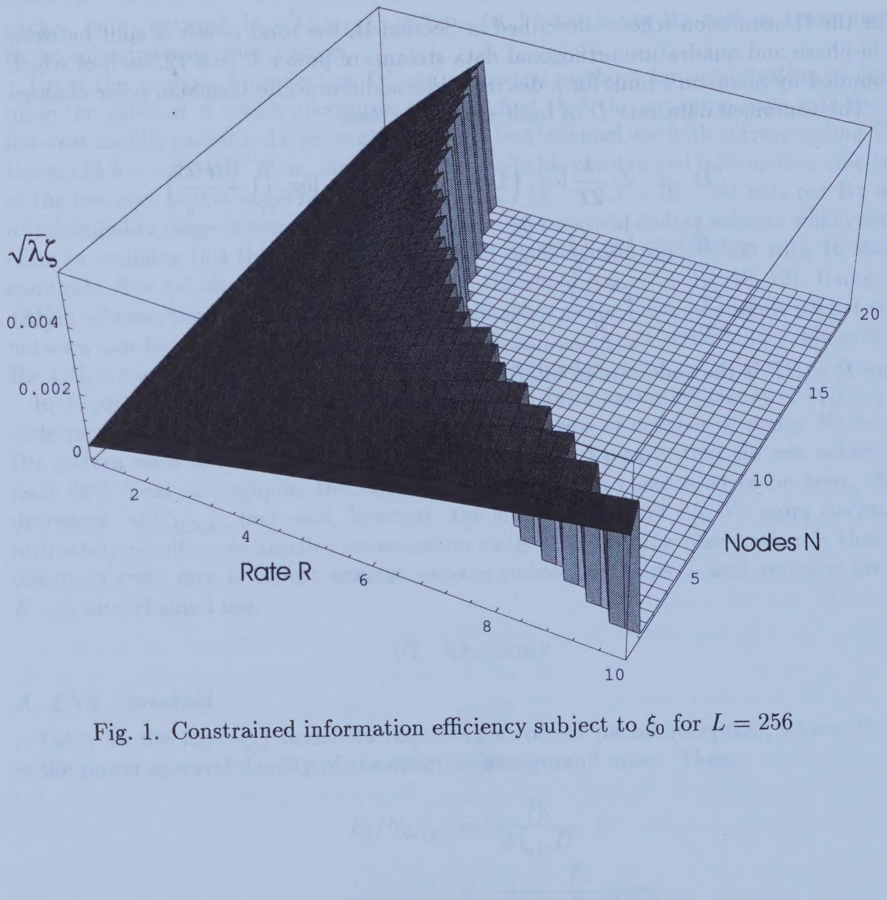


Fig. 1. Constrained information efficiency subject to ξ_0 for $L = 256$

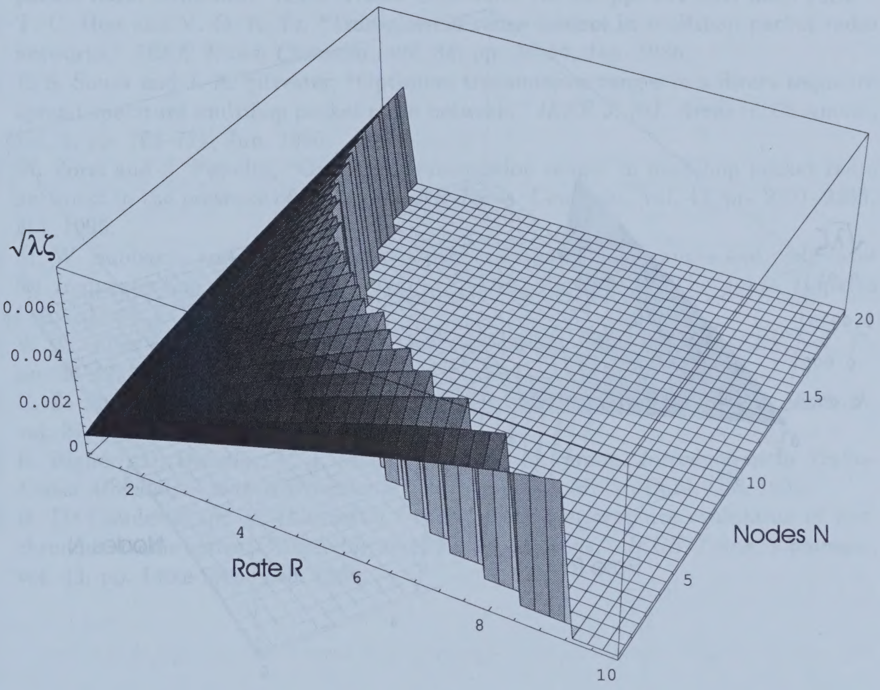


Fig. 2. Constrained information efficiency subject to ξ_0 for $L = 128$

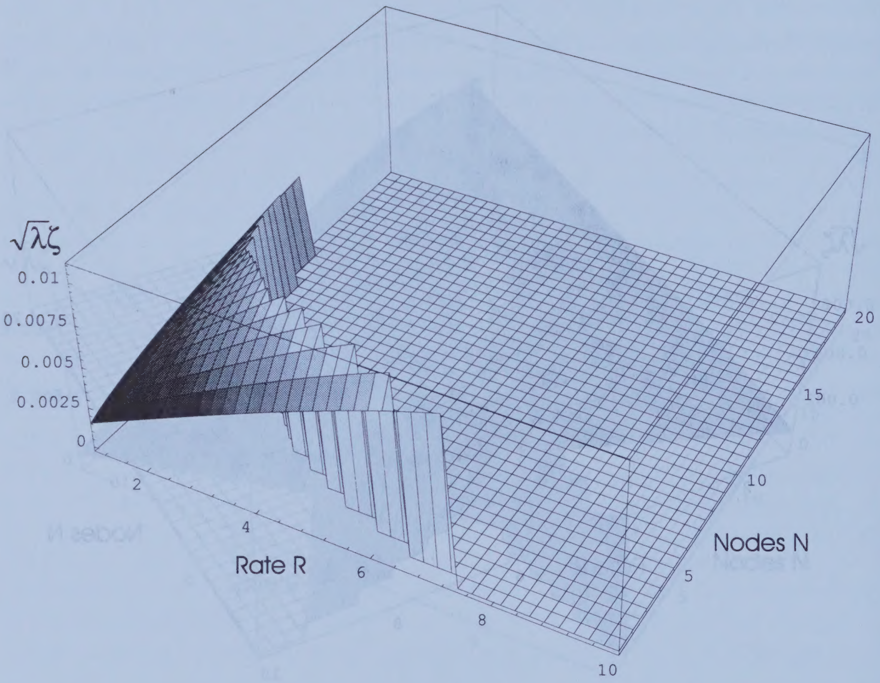


Fig. 3. Constrained information efficiency subject to ξ_0 for $L = 64$

REFERENCES

[1] L. Kleinrock and J. Silvester, "Spatial reuse in multihop packet radio networks," *Proc. IEEE*, vol. 75, pp. 156–167, Jan. 1987.

[2] B. Hajek, "Adaptive transmission strategies and routing in mobile radio networks," in *Conf. on Info. Sciences and Sys.*, pp. 373–378, Mar. 1983.

[3] L. Kleinrock and J. Silvester, "Optimum transmission radii for packet radio networks or why six is a magic number," in *Conf. Rec. Nat. Telecommun. Conf.*, pp. 4.3.1–4.3.5, Dec. 1978.

[4] J. A. Silvester, *On the spatial capacity of packet radio networks.* PhD thesis, UCLA, Dept. Computer Science, May 1980.

[5] H. Takagi and L. Kleinrock, "Optimal transmission ranges for randomly distributed packet radio terminals," *IEEE Trans. Commun.*, vol. 32, pp. 264–257, Mar. 1984.

[6] T. C. Hou and V. O. K. Li, "Transmission range control in multihop packet radio networks," *IEEE Trans. Commun.*, vol. 34, pp. 38–44, Jan. 1986.

[7] E. S. Sousa and J. A. Silvester, "Optimum transmission ranges in a direct-sequence spread-spectrum multihop packet radio network," *IEEE J. Sel. Areas in Commun.*, vol. 8, pp. 762–771, Jun. 1990.

[8] M. Zorzi and S. Pupolin, "Optimum transmission ranges in multihop packet radio networks in the presence of fading," *IEEE Trans. Commun.*, vol. 43, pp. 2201–2205, Jul. 1995.

[9] M. W. Subbarao and B. L. Hughes, "Optimum transmission ranges and code rates for frequency-hop packet radio networks," Tech. Rep. 97-02, The Johns Hopkins Univ., Dept. ECE, Mar. 1997.

[10] W. C. Fifer and F. J. Bruno, "The low-cost packet radio," *Proc. IEEE*, vol. 75, pp. 33–42, Jan. 1987.

[11] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–657, 1943.

[12] E. Biglieri, D. Divsalar, P. J. McLane, and M. K. Simon, *Introduction to Trellis-Coded Modulation with Applications.* New York: MacMillan Publ. Co., 1991.

[13] R. De Gaudenzi and F. Giannetti, "Analysis and performance evaluation of synchronous trellis-coded CDMA for satellite applications," *IEEE Trans. Commun.*, vol. 43, pp. 1400–1408, Feb. 1995.

Robust Adaptive Coded Modulation for Time-Varying Channels with Delayed Feedback

Dennis L. Goeckel
University of Massachusetts - Amherst
ECE Department
Box 35110
Amherst, MA 01003-5110
e-mail: goeckel@ecs.umass.edu

Abstract

The idea of using knowledge of the current channel fading values to optimize the transmitted signal in wireless communication systems has attracted substantial research attention in recent years. However, the practicality of this adaptive signaling has been questioned due to the variation of the wireless channel over time, which results in a different channel at the time of data transmission than at the time of channel estimation. The goal of the work presented here is to characterize the effects of this channel variation on the adaptive signaling paradigm and use this characterization to design adaptive signaling schemes that are effective for the time-varying channel. The design of both uncoded adaptive quadrature amplitude modulation (QAM) and adaptive coded modulation is considered; using only a single outdated fading estimate when neither the Doppler frequency nor the exact shape of the autocorrelation function of the channel fading process is known, adaptive trellis coded modulation schemes are designed that can provide a significant increase in bandwidth efficiency over their non-adaptive counterparts.

1 Introduction

The goal of wireless communication systems is to match the throughput of their wired counterparts. However, the limited bandwidth available and time-varying multipath fading inherent to the wireless link make this a difficult problem. One possible method of significantly increasing the bandwidth efficiency of communications systems operating over multipath fading channels is to employ adaptive signaling schemes, which use knowledge of the current channel fading values to optimize the transmitted signal. This idea has been of significant research interest in recent years [1, 2, 3, 4, 5, 6]; however, the utility of these results has been questioned due to the time-varying nature of the wireless channel, which results in a different channel at the time of data transmission than at the time of channel estimation. Only Cavers [1] has considered the effects of the channel variation on system design, and, more importantly, there has been no consideration of robustness nor has the characterization of the channel variation been used to design codes that are effective for a time-varying channel.

Based on previous work, it may appear that accommodating an adaptive rate system is not worth the effort. It can be shown that transmitter knowledge of the current

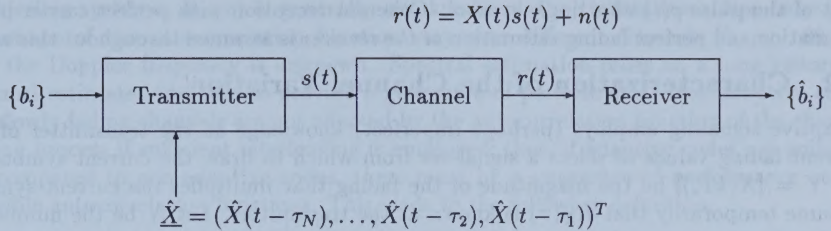

$r(t) = X(t)s(t) + n(t)$
$\{b_i\}$
Transmitter
$s(t)$
Channel
$r(t)$
Receiver
$\{\hat{b}_i\}$
$\underline{\hat{X}} = (\hat{X}(t-\tau_N), \ldots, \hat{X}(t-\tau_2), \hat{X}(t-\tau_1))^T$

Figure 1: System Diagram

channel fading values does not increase the channel capacity unless the transmit power is adapted [7], and even if the transmit power is adapted, the gain in channel capacity is small [5]. Furthermore, the recent establishment of trellis coded modulation that operates close to the cutoff rate [8] might imply that there is little gain available even at reasonable complexities. However, it will demonstrated in this paper that adaptive coding can provide significant gains relative to the codes presented in [8] for systems with low mobility, in particular indoor wireless systems in the 900 MHz ISM band with walking speed mobility.

In Section 2, a general framework for adaptive signaling over time-varying channels is introduced, and the effects of the channel variation are considered. The impact of the uncertainty in the statistics of the channel fading at the transmitter is also discussed, which leads to the definition of strongly robust signaling. The design of strongly robust uncoded adaptive quadrature amplitude modulation (QAM) schemes for the time-varying channel is presented in Section 3. In Section 4, a design method is presented for strongly robust adaptive trellis coded modulation operating over time-varying channels. Finally, Section 5 presents the conclusions.

2 System Model and Characterization

2.1 System Model

A baseband system model for the problem is shown in Figure 1. The independent and identically distributed (IID) sequence $\{b_i\}$ is the stream of information bits to be transmitted over the channel, and $\{\hat{b}_i\}$ is the corresponding stream of information bit estimates at the output of the receiver. Each information bit is assumed to be equally likely to be 0 or 1. The transmitted signal is given by $s(t) = \sum_{k=-\infty}^{\infty} z_k p(t - kT_s)$, where z_k is the k^{th} (complex) data symbol, $p(t)$ is a pulse shape that results in no intersymbol interference in the samples (spaced at T_s) of the output of the matched filter at the receiver, and $\frac{1}{T_s}$ is the symbol rate. The additive noise $n(t)$ is white Gaussian noise with two-sided power spectral density $\frac{N_0}{2}$. The fading is modeled as a complex multiplier $X(t)$, thus implying a frequency-nonselective channel appropriate for a narrowband wireless channel or a single subchannel of a wireless multicarrier system. The Gaussian wide-sense stationary uncorrelated scattering (GWSSUS) fading model is assumed, where the independent component Gaussian processes are zero mean with autocorrelation function $R_X(\tau)$; that is, $X(t) = X_R(t) + jX_I(t)$ is complex Gaussian, where $X_R(t)$ and $X_I(t)$ are the respective real and imaginary parts of $X(t)$, $j = \sqrt{-1}$, $E[X_R(t)] = E[X_I(t)] = 0$, and $E[X_R(t)X_R(t+\tau)] = E[X_I(t)X_I(t+\tau)] = R_X(\tau)$. It is also assumed that $X(t)$ varies slowly enough that analysis can be performed by assuming it is constant over the support of the pulse $p(t)$ of a single symbol. Coherent reception with perfect carrier phase estimation and perfect fading estimation *at the receiver* is assumed throughout this work.

2.2 Characterization of the Channel Variation

Adaptive signaling employs (perhaps imperfect) knowledge at the transmitter of the current fading values to select a signal set from which to draw the current symbol z_k. Let $Y = |X(kT_s)|$ be the magnitude of the fading that multiplies the current symbol. Assume temporarily that $R_X(\tau)$ is known at the transmitter. Let N be the number of outdated estimates employed, and define the $(N+1)$ by $(N+1)$ correlation matrix M_N as:

$$M_N = \begin{bmatrix} R_X(0) & R_X(\tau_N - \tau_{N-1}) & \dots & R_X(\tau_N - \tau_1) & R_X(\tau_N) \\ R_X(\tau_N - \tau_{N-1}) & R_X(0) & \dots & R_X(\tau_{N-1} - \tau_1) & R_X(\tau_{N-1}) \\ \dots & & & & \\ R_X(\tau_N - \tau_1) & R_X(\tau_{N-1} - \tau_1) & \dots & R_X(0) & R_X(\tau_1) \\ R_X(\tau_N) & R_X(\tau_{N-1}) & \dots & R_X(\tau_1) & R_X(0) \end{bmatrix}$$

Following the notation of [10, pg. 5], let the inverse of M_N be partitioned as

$$M_N^{-1} = \begin{bmatrix} Q^{-1} & \underline{r} \\ \underline{r}^T & p^{-1} \end{bmatrix}$$

where Q is an N by N matrix, $\underline{r}$ is a column vector of length N, and p is a scalar.

Since the focus here is on characterizing the effects of the channel variation independent of the channel estimation algorithm employed, assume that the outdated fading estimates are made perfectly; that is, $\hat{X}(kT_s - \tau_i) = X(kT_s - \tau_i), i = 1, \dots, N$. Since $X(kT_s)$ and the variables in $\hat{\underline{X}}$ are jointly Gaussian, $Y = |X(kT_s)|$ is Rician when conditioned on $\hat{\underline{X}}$. Thus, the conditional probability density function of Y is given by

$$p_{Y|\hat{\underline{X}}}(y|\underline{x}) = \frac{y}{\sigma^2} e^{-\frac{y^2+s^2}{2\sigma^2}} I_0\left(\frac{ys}{\sigma^2}\right) \qquad y \geq 0, \tag{1}$$

where $I_0(\cdot)$ is the 0th-order modified Bessel function, the noncentrality parameter is given by $s^2 = (p\underline{r}^T\underline{x}_R)^2 + (p\underline{r}^T\underline{x}_I)^2$, and $\sigma^2 = p$. Note that σ^2 is the mean squared error of a minimum mean squared error (MMSE) predictor of the current in-phase or quadrature fading value. For interpretation of (1), consider the $N = 1$ case. Defining $\rho = R_X(\tau_1)$, and normalizing such that $E[(X_R(kT_s))^2] = E[(X_I(kT_s))^2] = 1$ (which makes the average received energy twice the average transmitted energy and will be accounted for below), one obtains $s^2 = |\underline{x}|^2\rho^2$ and $\sigma^2 = 1 - \rho^2$. Define the specular-to-diffuse component (Rician) factor of (1) by $K = \frac{s^2}{2\sigma^2} = \frac{|\underline{x}|^2\rho^2}{2(1-\rho^2)}$. For a fixed correlation ρ, the effective channel conditioned on a single outdated fading estimate varies from a low signal-to-noise ratio (SNR) nearly Rayleigh fading channel to a high-SNR strongly Rician fading channel, with the type of fading depending on the value of the outdated estimate.

2.3 Robustness to Autocorrelation Function Uncertainties

If $R_X(\tau)$ is known exactly at the transmitter, using a large number of outdated fading channel estimates yields a linear MMSE predictor for the current in-phase and quadrature fading values that works well (i.e. $\sigma^2 \approx 0$) for systems with modest mobility; however, in wireless systems, $R_X(\tau)$ is generally not known at the transmitter. Not only does the shape of the autocorrelation function vary over different channels [9, pg. 88-89], but the Doppler frequency is unknown. Spectral estimation relies on a long history of channel estimates and cannot provide a guarantee of performance. Non-adaptive codes for slowly fading channels are not affected by the autocorrelation function of the channel fading process if sufficient interleaving is employed; thus, if adaptive codes are going to be compared to non-adaptive codes, there must be a guarantee of performance across possible autocorrelation functions. This leads to the following definition:

Definition: An adaptive signaling scheme is *strongly robust* with respect to $\mathcal{R}$ if it operates below the desired error probability for all autocorrelation functions $R_X(\tau) \in \mathcal{R}$.

In this paper, the design methods considered will guarantee strongly robust performance for each observation value $\underline{\hat{X}} = \underline{x}$, which guarantees strong robustness of the method. However, in general, this approach can greatly reduce the data rate of strongly robust adaptive signaling methods if more than a single outdated fading estimate is employed (i.e. $N > 1$). Thus, in this paper, the methods introduced will be applicable to two cases: strongly robust adaptive signaling with $N = 1$, and adaptive signaling with arbitrary N when $R_X(\tau)$ is known. Design for the second case is a simplification of design for the first case; thus, strongly robust adaptive signaling with a single outdated fading estimate will be considered. The $N = 1$ case is of considerable practical interest, in particular for bursty high-speed data applications, where only a single fading estimate is generally available from initial handshaking.

3 Adaptive Uncoded QAM

Per Section 2.3, the strongly robust $N = 1$ case will be considered, requiring performance to be guaranteed for all $\rho \in [\rho_{min}, 1]$, where ρ_{min} is the minimum value of $R_X(\tau_1)$.

3.1 Design Rules

Assume for illustrative purposes that the set of candidate signal sets are 0-QAM (no data transmitted), 2-QAM, 4-QAM, 16-QAM, 64-QAM with two-dimensional Gray mapping as considered in [6]. Temporarily, no energy adaptation is employed; in other words, the average energy of the signal constellation employed will be constant across symbols. Let P_b be the target bit error probability for the system, which operates at the average received signal-to-noise ratio $\frac{E_s}{N_0}$, where E_s is the average received energy per QAM symbol. Specification of the adaptive transmitter requires finding $\tilde{M}(h), \forall h$, where $\tilde{M}(h)$ is the number of signals in the QAM signal set employed when $|\hat{X}(kT_s - \tau_1)| = h$. Per Section 2.3, $\tilde{M}(h)$ is chosen such that P_b is maintained for each h:

$$\tilde{M}(h) = \max\{M : \sup_{\rho_{min} \le \rho \le 1} \tilde{P}_M\left(\frac{E_s}{N_0}, h, \rho\right) \le P_b\}, \tag{2}$$

where $\tilde{P}_M(\frac{E_s}{N_0}, h, \rho)$ is defined as the bit error probability of the M-QAM signal set at average received SNR $\frac{E_s}{N_0}$ when $R_X(\tau_1) = \rho$ and $|\hat{X}(kT_s - \tau_1)| = h$. Assume that maximum likelihood symbol detection given the current channel fading amplitude is employed on the samples of the matched filter output at the receiver. Conditioned on the value y of the current fading amplitude, the tight approximation [6]

$$P_M\left(y^2 \frac{E_s}{N_0}\right) \approx 0.2 \exp\left(-\frac{3}{4(M-1)} \frac{E_s}{N_0} y^2\right) \tag{3}$$

will be employed for all M for the design work for uncoded systems. Using (3) yields

$$\begin{aligned} \tilde{P}_M\left(\frac{E_s}{N_0}, h, \rho\right) &= E_Y\left[P_M\left(Y^2\frac{E_s}{N_0}\right) \middle| |\hat{X}(kT_s - \tau_1)| = h\right] && (4) \\ &\approx \begin{cases} \dfrac{0.2\exp\left[-\frac{h^2\rho^2}{2(1-\rho^2)}\left(1-\frac{1}{1+\frac{3}{2}\frac{E_s}{N_0}\frac{(1-\rho^2)}{(M-1)}}\right)\right]}{1+\frac{3}{2}\frac{E_s}{N_0}\frac{(1-\rho^2)}{(M-1)}} & \rho < 1 \\ 0.2\exp\left(-\frac{3}{4}\frac{E_s}{N_0}\frac{h^2}{(M-1)}\right) & \rho = 1 \end{cases} && (5) \end{aligned}$$

where the second line is obtained by substituting (1) and (3) into (4). From (2), (5) must be evaluated at its supremum on $\rho \in [\rho_{min}, 1]$. Since the right side of (5) is a continuous function on this closed interval, it achieves its maximum on this interval at a point which will be denoted ρ^*. The following solution is found by standard calculus techniques. Let

$$\tilde{\rho} = \begin{cases} 0 & h \geq \sqrt{2} \\ \sqrt{\left(1 + \frac{2(M-1)}{3}\frac{N_0}{E_s}\right)\frac{(2-h^2)}{2}} & 0 \leq h \leq \sqrt{2} \end{cases} .$$

The worst case correlation value is then given by

$$\rho^* = \begin{cases} \rho_{min} & \tilde{\rho} \leq \rho_{min} \\ \tilde{\rho} & \rho_{min} < \tilde{\rho} < 1 \\ 1 & 1 \leq \tilde{\rho} \end{cases} . \qquad (6)$$

The signal set is specified using (5) and (6) in $\tilde{M}(h) = \max\{M : \tilde{P}_M\left(\frac{E_s}{N_0}, h, \rho^*\right) \leq P_b\}$.

The rate of this method is limited by the following reason: with probability one, the estimate is more favorable than that required to use $\tilde{M}(h)$-QAM but not favorable enough to use $(\tilde{M}(h)+1)$-QAM. A solution to this problem is to employ energy adaptation. First, a signal set is chosen with no energy adaptation. Then (5) and (6) are used to decide the minimum energy required to maintain P_b given the channel estimate h; this method is similar to the power pruning presented in [11]. This minimum energy is employed for the current symbol and the remainder of the energy is saved for the next symbol.

3.2 Numerical Results

Figure 2 displays the probability of bit error of the scheme developed in Section 3.1 and the probability of bit error for a scheme that is designed based on a static channel (i.e. one that assumes $\rho = 1$). When designing for the static channel at $P_b = 10^{-5}$, the more conservative estimate $P_2(y^2\frac{E_s}{N_0}) \approx \frac{1}{2}\exp(-\frac{E_s}{2N_0}y^2)$ is employed, as the approximation of (3) is very poor at this point. Note that the probability of error targets are missed considerably by the scheme that assumes a static channel when there is even modest channel variability, particularly for low target bit error probabilities; this is due to (1) being a Rician density. Figure 3 plots the rate of the scheme of Section 3.1 versus signal-to-noise ratio for a number of values of ρ_{min} at the target bit error probability $P_b = 10^{-5}$.

4 Adaptive Trellis Coded Modulation

4.1 Paradigm for Adaptive Trellis Coded Modulation

Adaptive signaling is viewed as a powerful method for highly bandwidth efficient modulation, and thus the focus here is on adaptive trellis coded modulation as in [2, 3, 4]. The

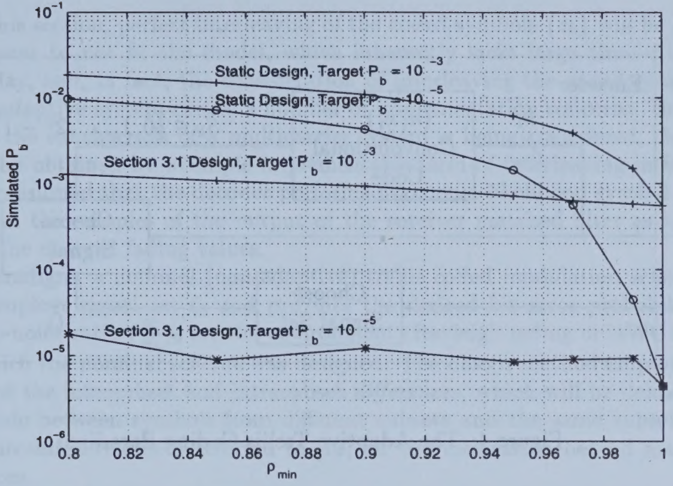


Figure 2: Simulated probability of bit error of adaptive uncoded QAM design schemes versus ρ_{min} at average received signal-to-noise ratio $\frac{E_s}{N_0} = 15$ dB per QAM symbol.

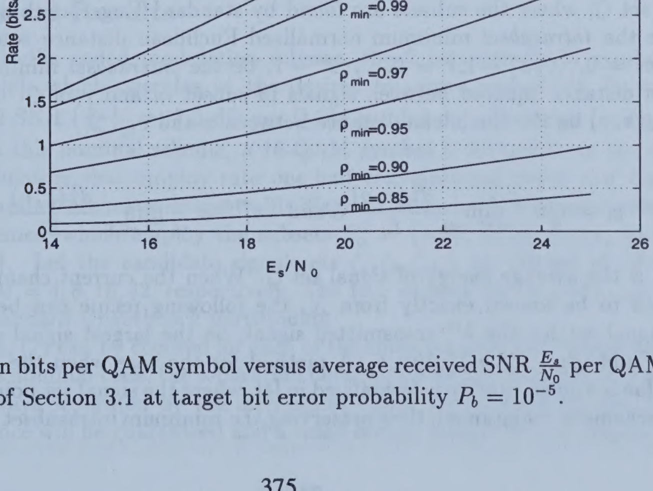


Figure 3: Rate in bits per QAM symbol versus average received SNR $\frac{E_s}{N_0}$ per QAM symbol for the scheme of Section 3.1 at target bit error probability $P_b = 10^{-5}$.

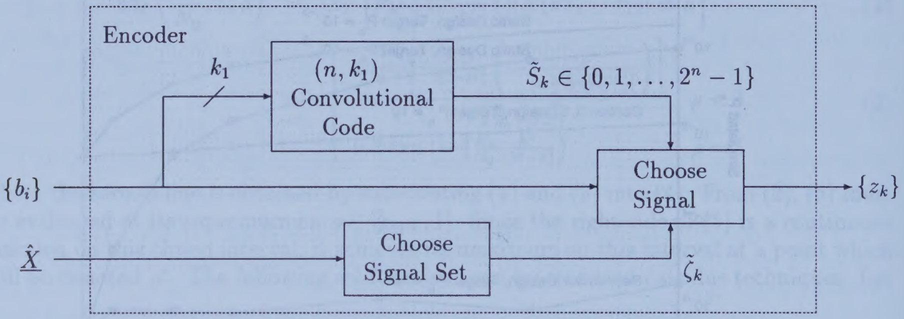


Figure 4: The Adaptive Trellis Coding Paradigm

adaptive trellis coded modulation paradigm that will be followed is shown in Figure 4. For this paradigm, the number of subsets 2^n is kept constant across all signal sets; only the number of signals per subset is adapted. The current symbol z_k is chosen as follows: k_1 bits are input to the convolutional coder, which produces n bits that are used as the subset selector $\tilde{S}_k$. The signal constellation $\tilde{\zeta}_k$ (and thus the subsets) is chosen from the allowable signal sets $\zeta_1, \zeta_2, \ldots, \zeta_L$ (with possibly scaling for energy adaptation) based on the information about the current fading provided by $\underline{\hat{X}}$. Finally, the appropriate number of data bits (base two logarithm of the size of a subset of $\tilde{\zeta}_k$) are used to choose z_k from subset $\tilde{S}_k$ of $\tilde{\zeta}_k$. There are two design considerations that must be considered: how to design the trellis structure and subset mapping, and how to choose the signal set $\tilde{\zeta}_k$ based on $\underline{\hat{X}}$. The first item is solved well in the literature for non-adaptive coding; thus, it will be assumed here that there exists a nominal (non-adaptive) trellis coded modulation scheme with signal set ζ_0 that achieves the desired bit error probability P_b. The same trellis structure and subset mapping of this nominal scheme will be employed.

Consider the choice of the signal set $\tilde{\zeta}_k$. Let $S_l^r, r = 0, \ldots, 2^n - 1$ be the r^{th} subset of signal set ζ_l, where the subsets are found by standard Ungerboeck partitioning [12]. Let g_l be the *intrasubset* minimum normalized Euclidean distance squared of ζ_l, and let $u_l^{m,r}, m = 0, \ldots, 2^n - 1, r = 0, \ldots, 2^n - 1$, be the *intersubset* minimum normalized Euclidean distance squared between signals in subset m and r of ζ_l. In other words, letting $d_E(x, v)$ be the Euclidean distance between x and v,

$$g_l = \min_{S_l^r} \min_{x, v \in S_l^r, x \neq v} \frac{d_E^2(x, v)}{E_l} \quad \text{and} \quad u_l^{m,r} = \min_{x \in S_l^m, v \in S_l^r} \frac{d_E^2(x, v)}{E_l},$$

where E_l is the average energy of signal set ζ_l. When the current channel fading value is assumed to be known exactly from $\underline{\hat{X}}$, the following recipe can be employed: let $\tilde{\zeta}_k$, the signal set for the k^{th} transmitted signal, be the largest signal set ζ_l such that $y^2 g_l \geq g_0$ and $y^2 u_l^{m,r} \geq u_0^{m,r}, \forall m, r$. A method for the case when the current channel fading value is known exactly is formalized in [4], where the signal set point density of the nominal scheme is maintained, thus preserving the minimum intrasubset and intersubset distances.

4.2 Design Rules

Throughout this section, perfect interleaving of the coded symbols $\{z_k\}$ will be assumed. This may appear to not fit the model, which inherently lacks large time diversity at reasonable delay, and, in fact, the consideration of interleaving the coded symbols in a single carrier adaptive system causes a number of implementation concerns. However, if the coding of the subchannels of a multicarrier system is being considered, the desired diversity can be obtained by a simple implementable form of interleaving in frequency. It will also be assumed that the decoder performs maximum likelihood symbol sequence estimation with the samples of the output of the receiver matched filter as input and knowledge of the channel fading values.

Per the paradigm, a nominal (non-adaptive) trellis coded modulation scheme is assumed that employs signal set ζ_0 and provides the desired bit error probability P_b at some signal-to-noise ratio $(\frac{E_s}{N_0})_0$ for the environment (Rayleigh fading or AWGN, for example) for which the nominal scheme was designed. The design method employed is the maintenance of the intersubset and intrasubset *differences*, which will be defined as the ability to decide between symbols from different subsets and the same subset, respectively. The interested reader is referred to [13] for the motivation behind maintaining these differences.

To maintain all of the minimum intersubset differences, the design must be such that:

$$\sup_{\rho \geq \rho_{min}} \int_0^\infty e^{-y^2 u_l^{m,r} \frac{E_s}{8N_0}} p_{Y|h}(y|h)dy \leq D_0(0, u_0^{m,r} E_s) \tag{7}$$

for all $m = 0, \ldots, 2^n-1$ and $r = 0, \ldots, 2^n-1$, $m \neq r$, where $D_0(w,v)$ is the Bhattacharyya bound on the error probability of choosing w when v is correct for the nominal code. To maintain the intrasubset differences, the design is made such that:

$$\sup_{\rho \geq \rho_{min}} \frac{1}{2} \int_0^\infty e^{-y^2 g_l \frac{E_s}{8N_0}} p_{Y|h}(y|h)dy \leq P_b. \tag{8}$$

The form of the left sides of (7) and (8) is identical to equations of Section 3.1, and thus their manipulation is omitted.

4.3 Numerical Results

For a nominal trellis coded modulation scheme, the 2 bits per QAM symbol 8-state I-Q code from [8], which operates at the desired bit error probability $P_b = 10^{-5}$ at average received SNR $\left(\frac{E_s}{N_0}\right)_0 = 17.5$ dB per QAM symbol on a Rayleigh fading channel, is employed. In this nominal scheme, a 16-QAM symbol is formed from the output of two coded modulators that employ rate one-half convolutional codes and 4-ary ASK. There are no parallel branches in the trellis diagram of each of these component coded 4-ary ASK schemes, which employ the subsets $S_0^0 = \{-3\}$, $S_0^1 = \{-1\}$, $S_0^2 = \{+1\}$, and $S_0^3 = \{+3\}$. Let the candidate signal sets $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ be the set of M-ary ASK constellations, $M = 4, 8, 16, 32$, respectively. The four subsets for each of $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ are defined using standard Ungerboeck partitioning.

For a given h, (7) can be checked for each of the three possible intersubset differences; however, for this nominal code, the separation between trellis paths leaving and remerging at a node is always separated by the intersubset difference between S_0^0 and S_0^2. Thus, only this difference will be guaranteed and a small energy margin will be employed; it has

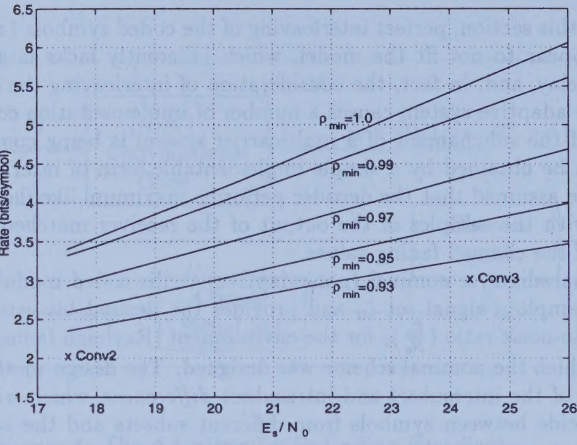


Figure 5: Simulated rate of adaptive 8-state trellis coded modulation in bits per QAM symbol as a function of the average received SNR $\frac{E_s}{N_0}$ per QAM Symbol. "Conv2" and "Conv3" correspond to 8-state I-Q Codes from [8] that operate at bit error probability 10^{-5}. The simulated P_b for each data point is less than the target 10^{-5}.

been found that this results in a better scheme than guaranteeing all of the intersubset differences when the trellis of the nominal scheme has no parallel branches. Making this intersubset guarantee for each l yields h_l^{tr}, the threshold such that for $h \geq h_l^{tr}$, ζ_l can be employed according to the intersubset differences. Likewise, (8) can be used to find h_l^{par}, the threshold such that for $h \geq h_l^{par}$, ζ_l can be employed according to the intrasubset differences. The threshold for h above which ζ_l can be used is then given by $\max(h_l^{tr}, h_l^{par})$. Energy adaptation will be employed similarly to the method for uncoded systems; the minimum energy that is required to guarantee (7) and (8) is used for this symbol; the remainder of the energy is saved for the next symbol. Numerical results for the rate of this adaptive trellis coded modulation scheme are shown in Figure 5. These results indicate that there is the possibility of significant bandwidth efficiency increases by employing adaptive coding. For example, in the 900 MHz indoor environment, where a suitable class of autocorrelation functions might be all functions in $\{R_X(\tau) : R_X(\tau) \geq J_0(2\pi 6\tau)\}$ that satisfy the properties of an autocorrelation function [9], a large throughput gain is possible if data transmission follows channel estimation by 5 to 10 ms, a time during which a number of packets can be transmitted in a high-speed wireless data system.

5 Conclusions

Although the theory of adaptive coding for fading channels suggests large throughput gains over non-adaptive coding, it has been avoided in practice because of perceived problems with the variation of the fading channel between channel estimation and data transmission. In this paper, it has been shown that these perceptions are well-founded: the delay between channel estimation and data transmission greatly alters the nature of the adaptive signaling problem. However, it has also been demonstrated that by designing with this variability in mind, the system designer can specify adaptive coding schemes that display large throughput gains over conventional schemes for systems with low mobility. This was done without requiring knowledge of either the Doppler frequency or the exact autocorrelation shape of the channel fading process.

References

[1] J. Cavers, "Variable-Rate Transmission for Rayleigh Fading Channels," *IEEE Transactions on Communications,* Vol. 20, pp. 15-22, February 1972.

[2] J. Bingham, "Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come," *IEEE Communications Magazine,* pp. 5-14, May 1990.

[3] S. Alamouti and S. Kallel, "Adaptive Trellis-Coded Multiple-Phase-Shift Keying for Rayleigh Fading Channels," *IEEE Transactions on Communications,* Vol. 42, pp. 2305-2314, June 1994.

[4] A. Goldsmith, "Variable-Rate Coded MQAM for Fading Channels," *Proceedings of the IEEE Global Communications Conference - Communications Theory Miniconference,* pp. 186-190, 1994.

[5] A. Goldsmith, "Capacity and Dynamic Resource Allocation in Broadcast Fading Channels," *Proceedings of the Allerton Conference on Communications, Control, and Computing,* pp. 915-924, 1995.

[6] S. Chua and A. Goldsmith, "Variable-Rate Variable-Power MQAM for Fading Channels," *Proceedings of the 46th IEEE Conference on Vehicular Technology,* pp. 815-819, May 1996.

[7] T. Ericson, "A Gaussian Channel with Slow Fading," *IEEE Transactions on Information Theory,* pp. 353-355, May 1970.

[8] S. Al-Semari and T. Fuja, "I-Q TCM: Reliable Communication over the Rayleigh Fading Channel Close to the Cutoff Rate," *IEEE Transactions on Information Theory,* Vol. 43, pp. 250-262, January 1997.

[9] K. Pahlavan and A. Levesque, *Wireless Information Networks,* New York: John Wiley & Sons, 1995.

[10] K. Miller, *Multidimensional Gaussian Distributions,* New York: John Wiley & Sons, 1964.

[11] E. Feig, "Practical Aspects of DFT-based Frequency Division Multiplexing for Data Transmission," *IEEE Transactions on Communications,* Vol. 38, pp. 929-932, July 1990.

[12] G. Ungerboeck, "Channel Coding with Multilevel/Phase Signals," *IEEE Transactions on Information Theory,* Vol. 28, pp. 55-67, January 1982.

[13] D. Goeckel, "Adaptive Coding Using Outdated Fading Estimates," submitted to the *IEEE Transactions on Communications,* August 1997.

Adaptive Channel Coding in DS-CDMA Networks Under Imperfect State Information

Seong-Jun Oh and Kimberly M. Wasserman *
EECS Department
University of Michigan
1301 Beal Avenue, Ann Arbor, MI 48109-2122

Abstract

In this short paper, we present some preliminary results regarding adaptive channel coding in direct sequence spread spectrum networks. The network consists of two mobile users and one base station. One of the users is a data source and the other an "on/off" voice source. The packets transmitted by the data source are corrupted by additive white Gaussian noise and, if the voice user is on, multiple access interference. The data source receives acknowledgements from the base station merely indicating whether its packet transmissions were correctly received. The data source must use these acknowledgements, together with knowledge of the statistics of the voice source to adaptively control its channel code rate so that throughput is maximized. We show the optimal policy is one that periodically compares a suitable sufficient statistic with a threshold.

1 Introduction

Consider a direct-sequence code division multiple access (DS-CDMA) network consisting of a single base station and two mobile users. The users transmit over a common broadband channel to the base station using different spreading codes (i.e., PN or "chipping" sequences). The spreading codes have non-zero cross correlation; therefore the signal power generated by one user contributes as multiple access interference to the signal transmitted by the other user.

One of the two users is a data source and the other is an "on/off" voice source; the statistical behavior of the voice source is known to the data source. If the state of the voice source is "on", a packet transmitted by the data source is corrupted by both multiple access interference *and* additive white Gaussian noise; otherwise, a transmission is corrupted by only additive white Gaussian noise.

Both users apply open-loop power control, via time division duplexing (TDD) between the forward and reverse link to perfectly maintain a prespecified received signal power level at the base station. The data source employs forward error correction (FEC) and is allowed to control the packet error rate it experiences by adapting the channel code rate. The optimal value depends on the state of the voice source; however, the only information available to the source is the stream of acknowledgements it receives from the base station, which merely indicate whether or not the packets it transmits are correctly received.

We are interested in obtaining the adaptive channel code rate policy that achieves maximal throughput.

---

*This research was supported by the U.S. Army Research Office under grant DAAH04-96-1-0177.

2 Network Model

The time-axis is divided into contiguous equal-length time slots. The time duration of a slot is equal to the time duration of a packet. We assume each packet is an n-bit codeword of an (n, k) linear block code, where $n - k$ is the number of redundant or parity bits and k is the number of information bits.

The data source has an infinite number of information bits, and transmits a packet in each time slot. At the beginning of a time slot, the data source must decide how many of the n bits of the packet to be transmitted will be parity bits and how many will be information bits, that is, the source must decide on a channel code rate. The optimal value depends on whether or not the voice source will be active (i.e., in the "on" state) during the time slot. Let $n - k_1$ be the optimal number of parity bits given the voice user is active, and $n - k_0$ the optimal number given the voice user is inactive, where $k_0 > k_1$.

Define the random process $V = \{V_t; t \in \mathbb{Z}_+\}$, where $V_t \in \{0, 1\}$ indicates the state of the voice source during the t^{th} time slot. Specifically, the voice source will be active during $[t, t+1)$ if $V_t = 1$, and silent otherwise, $t \in \mathbb{Z}_+$. The process V is a two-state time-homogeneous and stationary Markov chain with transition probabilities given by $P\{V_t = 1 \mid V_t = 0\} = \lambda$ and $P\{V_t = 0 \mid V_t = 1\} = \mu$. Also define the random sequence $A = \{A_t; t \in \mathbb{N}\}$, where $A_t \in \{0, 1\}$ is the acknowledgement (feedback) transmitted to the data source from the base station at the end of the $(t - 1)^{st}$ time slot. Specifically, the packet transmitted by the data source during $[t - 1, t)$ is received correctly if and only if $A_t = 1$, $t \in \mathbb{N}_+$. We assume that packets which are not correctly received are lost and not retransmitted. Define the control sequence $U = \{U_t; t \in \mathbb{Z}_+\}$ where $U_t \in \{k_0, k_1\}$ indicates that $n - U_t$ bits of the packet transmitted by the source during $[t, t + 1)$ are redundant.

Define $P_t = P\{V_t = 1 \mid A_t = a_t, \ldots, A_1 = a_1, U_{t-1} = u_{t-1}, \ldots, U_0 = u_0\}$ to be the posterior probability that the voice source is in the "on" state during $[t, t + 1)$ given the history of the sequence of controls and acknowledgements, $t \in \mathbb{Z}_+$. The probability that a packet is successfully transmitted depends on both the state of the voice source and the number of parity bits. Let $p_u(a) = P\{A_t = a \mid V_t = 0, U_t = u\}$ and $q_u(a) = P\{A_t = a \mid V_t = 1, U_t = u\}$, $a \in \{0, 1\}$, $u \in \{k_0, k_1\}$, and $t \in \mathbb{Z}_+$.

Define the random sequence $R = \{R_t; t \in \mathbb{Z}_+\}$, where $R_t = P_t/(1 - P_t)$ are the odds in favor of the event $\{V_t = 1\}$, $t \in \mathbb{Z}_+$. It can be shown that

$$R_t = L_{U_{t-1}}(A_t) \left(\frac{\lambda + (1 - \mu)R_{t-1}}{1 - \lambda + \mu R_{t-1}} \right), \tag{1}$$

where $L_u(a) = q_u(a)/p_u(a)$ is a likelihood ratio, $a \in \{0, 1\}$, $u \in \{k_0, k_1\}$, and $t \in \mathbb{Z}_+$. Therefore R_t can be computed recursively starting with $R_0 = P\{V_t = 1\}/(1 - P\{V_t = 1\})$, $t \in \mathbb{Z}_+$.

3 Optimal Adaptive Channel Coding

We are interested in obtaining the channel coding policy that maximizes the average number of successfully transmitted bits per slot (i.e., the number of error-free bits received per slot or "goodput"). Towards this, consider the total expected discounted cost function

$$J(r) = E\left[\sum_{t=0}^{\infty} \alpha^t C(R_t, U_t) \middle| R_0 = r \right], \tag{2}$$

where $0 < \alpha < 1$ and

$$C(r, k_i) = \begin{cases} \frac{r}{1+r} \frac{n-k_0}{n} & \text{if } i = 0 \\ \frac{1}{1+r} \frac{n-k_1}{n} & \text{if } i = 1, \end{cases} \tag{3}$$

$r \in \mathbb{R}_+$. Given the odds r in favor of an active voice source, the cost $C(r, k_0)$ ($C(r, k_1)$) is the product of the "redundancy rate" and and the posterior probability that the voice source is in the "on" ("off") state, and can be thought of as the throughput penalty to be paid by transmitting the packet with either too many parity bits or too few, that is, transmitting with too low or too high of a code rate, $r \in \mathbb{R}_+$, $i \in \{0, 1\}$.

The following proposition characterizes the structure of the optimal adaptive channel coding policy, and is proven via dynamic programming techniques[1].

Proposition 1 *There exists a real number r^* such that, when $R_t > r^*$, the optimal policy chooses $U_t = k_1$ (lower code rate), and when $R_t \leq r^*$, $U_t = k_0$ (higher code rate).*

References

[1] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, NJ, 1987.

Optimal Power Adaptation in Digital Repeaters

Dennis L. Goeckel
University of Massachusetts - Amherst
ECE Department
Amherst, MA 01003-5110
e-mail: goeckel@ecs.umass.edu

Abstract

The optimal allocation of power at the relay in a two-channel cascade as a function of the reliability of the decoded bits is considered. For systems where both links are uncoded, very minor gains are obtained by this technique, which essentially couples errors between the two links to reduce the overall error probability of the cascade. For systems where soft decision decoding of a convolutional code is performed at the receiver of the second channel, the gains can be more substantial but are due partially to erasure insertion at the relay.

1 Introduction

Motivated by the regenerative satellite problem, the classical problem of communication over two cascaded channels [1, 2, 3] to minimize the probability of error of the overall cascade is reopened. It was recognized in early works ([1] as referenced in [2]) that proper processing at the relay is necessary for optimizing data transmission over the cascade. Throughout this short paper, it will be assumed that the complexity available at the relay between the two channels (see Figure 1) is significantly less than that at the transmitter and receiver; in fact, the relay will be restricted to processing one symbol at a time. From Metzner [3], two statements of interest can be inferred:

1. If the relay is peak-power limited and the receiver is making hard decisions, the optimal relay is the standard repeater, which simply remodulates the most likely symbol transmitted.
2. If the relay is peak-power limited and the receiver is using soft information, the optimal relay may not be the standard repeater.

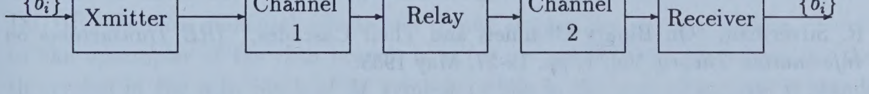


Figure 1: Block Diagram of a Cascade of Two Channels

In this paper, we consider each of these two cases under an average power constraint for the repeater.

2 Uncoded Links

Let E_s be the average symbol energy on the second link and $\underline{R}$ be a vector of reliability information (for example, the magnitude of a matched filter output) for decoded bits from the first link. The goal is to find the function $\gamma(\underline{r})$, where $\gamma(\underline{r})E_s$ is the energy employed on the second link when $\underline{R} = \underline{r}$, that minimizes the overall error probability

$$P(E) = \int_0^\infty \left(P(E_1|\underline{r}) + P(E_2|\underline{r}) - 2P(E_1|\underline{r})P(E_2|\underline{r})\right) p_{\underline{R}}(\underline{r})d\underline{r}$$

given the constraint $\int_0^\infty \gamma(\underline{r})p_{\underline{R}}(\underline{r})d\underline{r} = 1$, where E_1 corresponds to an error on the first link and E_2 to an error on the second link. This problem can be solved by employing Lagrange multipliers. When binary modulation is employed when both channels are additive white Gaussian noise (AWGN) channels, numerical results indicate that there is only a small performance gain by employing power adaptation. When the channels are assumed instead to be Rayleigh fading channels and the repeater performs power adaptation based on both the fading value and the magnitude of the matched filter output from the first channel, a gain of approximately 0.2-0.3 dB over the standard repeater is observed for the error rate range from 0.2 to 0.03 when the signal to noise ratios of the two channels are identical.

3 Soft Output Systems

In this section, all modulation will be assumed to be coherently detected binary phase-shift keying and both channels will be AWGN channels. A rate $\frac{1}{2}$, constraint length 5 convolutional code is employed in the system. The convolutional code will be decoded at the receiver using the Viterbi algorithm, which is *not* guaranteed to be the maximum likelihood sequence estimator if processing is done at the repeater. The optimal power adaptation for the repeater is difficult to find. Instead, the adaptive repeater either employs no energy (similar to "erasure insertion", although the receiver does not know where the erasures are located), or remodulates the most likely transmitted signal at a power level such that the overall average power is correct. The erasure insertion region is found numerically by optimizing the performance of a soft-decision decoded repetition code of length equal to the minimum distance of the convolutional code. This type of power adaptation leads to a gain of 0.2-0.5 dB over the bent pipe and standard repeaters for error ranges from 10^{-3} to 10^{-4} when the signal-to-noise ratios of the two channels are identical.

Acknowledgment: The assistance of Russ Nelson in the generation of numerical results is gratefully acknowledged.

References

[1] C. DeSoer, "Communication through Channels in Cascade," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1953.

[2] R. Silverman, "On Binary Channels and Their Cascades," *IRE Transactions on Information Theory,* Vol. 1, pp. 19-27, May 1955.

[3] J. Metzner, "Binary Relay Communications and Decision Feedback," *1959 IRE National Convention Record,* Part 4, pp. 112-122, 1959.

REDUNDANT FILTERBANK PRECODERS AND EQUALIZERS: OPTIMAL DESIGNS AND BLIND SIGNAL RECOVERY *

Anna Scaglione[1], *Georgios B. Giannakis*[2], *Sergio Barbarossa*[1]

[1]Infocom Dept., Univ. of Rome "La Sapienza", via Eudossiana 18, 00184 Roma, ITALY
[2] Dept. of Electrical Engr., Univ. of Virginia, Charlottesville, VA 22903-2442, USA

Abstract: *Transmitter redundancy introduced using filterbank precoders generalizes existing modulations including OFDM, DMT, TDMA, and CDMA schemes encountered with single- and multi-user communications. Sufficient conditions are derived to guarantee that with FIR filterbank precoders FIR channels are equalized perfectly in the absence of noise by FIR zero-forcing equalizer filterbanks, irrespective of the channel zero locations. Multicarrier transmissions through frequency-selective channels can thus be recovered even when deep fades are present. Jointly optimal transmitter-receiver filterbanks are also designed based on maximum output SNR and minimum mean-square error criteria under zero-forcing and fixed transmitted power constraints. More important, direct blind FIR equalizer filterbanks become possible which are: (i)deterministic, zero-forcing and capable of equalizing any FIR channel with small data sets; (ii)simple computationally and consistent statistically; (iii)equipped with rich structure allowing blind block synchronization and online algorithms.*

1 Introduction and Problem statement

Redundancy at the transmitter builds diversity in the input of digital communication systems and is well motivated for designing error correcting codes; see e.g., [1] and [4] for block-by-block transmissions. But recently, input diversity has been exploited for ISI suppression using filterbank precoders operating in the complex (as opposed to Galois) field, (see e.g., [3, 11]). In [3], filterbank precoders and blind equalizers were proposed in order to: (i)improve the rate reduction induced by the repetition precoders of [9]; and (ii)obviate channel zero restrictions imposed by spatio-temporal output diversity methods relying upon fractional sampling and/or multiple-antenna reception [10]. Input diversity and deep fades due to the channel nulls are also encountered with orthogonal frequency division multiplexing (OFDM) [2], discrete-multitone modulation (DMT) [5], and pseudo-random or wavelet filterbank precoders used in code division or discrete-wavelet multiplexing (CDMA/DWMA) [8], [6]. The filterbank framework considered herein encompasses these modulations and equalization schemes as special cases [7].

Consider the discrete-time multirate equivalent model of the baseband communication system in Fig. 1. Downsamplers and upsamplers perform blocking (i.e., multiplexing) and un-blocking (de-multiplexing) operations. The ratio $(P-M)/P$ represents the amount of redundancy introduced. Transmit filters $\{f_m(n)\}_{m=0}^{M-1}$ and receive-filters $\{g_p(n)\}_{p=0}^{P-1}$ are FIR of maximum order $P-1$ and K respectively, while the Lth order channel $\{h(l)\}_{l=0}^{L}$ includes multipath effects and timing ambiguities as delay factors. The input to the upsampler of the mth branch is $s_m(n) := s(nM+m)$. It represents the m-th symbol in the n-th block of M symbols, while in the multi-user case it stands for

---

*The work in this paper was supported by NSF-MIP grant no. 9424305.

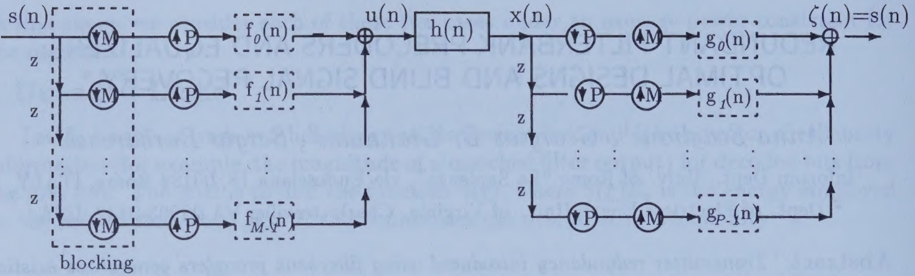


Figure 1: - Multirate discrete-time transmitter/channel/receiver model

the mth user's bits. With the insertion of $P-1$ zeros, the corresponding upsampler's output is: $u_m(n) = \sum_i s_m(i)\delta(n-iP)$, where $\delta(n)$ denotes Kronecker's delta. We will assume Nyquist signaling pulses; hence, their effect disappears in discrete-time and the transmitted sequence is: $u(n) = \sum_{m=0}^{M-1} u_m(n) = \sum_i \sum_{m=0}^{M-1} s_m(i) f_m(n-iP)$. The received samples, $y(n) = x(n) + v(n)$, consist of the noise-free data:

$$x(n) = \sum_{i=-\infty}^{+\infty} \sum_{m=0}^{M-1} s_m(i) \sum_{l=0}^{L} h(l) f_m(n-l-iP), \tag{1}$$

plus additive zero-mean stationary noise with covariance matrix $\boldsymbol{R}_{vv}$ ($= \sigma_{vv}^2 \boldsymbol{I}$ when the noise is white). Thanks to the FIR nature of $h(n)$ and $f_m(n)$, it can be shown [1] that for $M > L$ the polyphase components $\{x_p(n)\}_{p=0}^{P-1} := \{x(nP+p)\}_{p=0}^{P-1}$ of $x(n)$ are:

$$x_p(n) = \sum_{m=0}^{M-1} s_m(n) \sum_{l=0}^{L} h(l) f_m(p-l) + \sum_{m=0}^{M-1} s_m(n-1) \sum_{l=0}^{L} h(l) f_m(P+p-l)\,, \tag{2}$$

where the terms $s_m(n)$ and $s_m(n-1)$ imply that the Lth order ISI induced by the channel entails at most two consecutive M data blocks.

Let us define $M \times 1$ vector $\boldsymbol{s}(n) := (s_0(nM) s_1(nM) \cdots s_{M-1}(nM))^T$, $P \times 1$ vectors $\boldsymbol{x}(n) := (x(nP) x(nP+1) \cdots x(nP+P-1))^T$, $\boldsymbol{f}_m := (f_m(0) \cdots f_m(P-1))^T$, $P \times M$ precoder matrix $\boldsymbol{F} := (\boldsymbol{f}_0 \cdots \boldsymbol{f}_{M-1})$, and $P \times P$ Toeplitz lower (upper) triangular matrix $\boldsymbol{H}_0$ ($\boldsymbol{H}_1$) with first (last) column $(h(0) \cdots h(L) 0 \cdots 0)^T$ $((h(1) \cdots h(L) 0 \cdots 0)^T)$.

Based on these definitions, we can cast (2) in matrix form:

$$\boldsymbol{x}(n) = \boldsymbol{H}_0 \boldsymbol{F} \boldsymbol{s}(n) + \boldsymbol{H}_1 \boldsymbol{F} \boldsymbol{s}(n-1), \tag{3}$$

and state this paper's operational assumptions as follows:
(a0) Channel $h(l)$ is Lth order FIR with $h(0), h(L) \neq 0$.
(a1) (P, M) are chosen such that the triplet (P, M, L) satisfies: $P > M > L$
(a2) Filters $\{f_m(n)\}_{m=0}^{M-1}$ are linearly independent; i.e., $rank(\boldsymbol{F}) = M$, which guarantees one-to-one mapping and thus recovery of $s(n)$ from the coded symbols $u(n)$.

Under (a0)-(a2), our objective is threefold: (i)to develop sufficient conditions for existence of FIR zero-forcing (ZF) filterbanks which in the absence of noise equalize perfectly FIR channels using FIR precoder filterbanks (Section 2); (ii)to derive jointly optimal FIR transmitter-receiver filterbank pairs which in the presence of noise maximize the output SNR, or minimize mean-square error under transmitted power constraint (Section 3); and (iii)to develop blind channel estimation, block synchronization, and direct equalization schemes by exploiting the input redundancy (Section 4).

---

[1] Simulations and proofs of claims and theorems in this paper can be found in the journal papers [7];

2 FIR-ZF Equalizing Filterbanks

With moderate or large number of filters M in the precoder, the maximum likehood receiver implemented with Viterbi's algorithm has prohibitively large complexity which motivates looking for linear (and preferably low order FIR) equalizing filterbanks. In this section we will focus on ZF solutions because they offer (almost) perfect symbol recovery in (high SNR) noise free environments and their performance in terms of error probability is easily computable. Our first result is summarized in the next theorem:

Theorem 1(leading equalizer zeros): *Assume* (**a0**)-(**a2**), *and in addition:*

(**a3**)*Each column of the* $QP \times QP$ *matrix* $(\boldsymbol{I}_{Q\times Q} \otimes \boldsymbol{F})$ *cannot be expressed as a linear combination of less than* $L+1$ *Vandermonde vectors.*

(**a4**) *For a given* L, *the triplet* (P, M, Q) *is selected to satisfy* (**a1**) *and* $P \geq M + \lceil L/Q \rceil$. *Then, for a given precoder* $\boldsymbol{F}$ *and channel matrices* $\boldsymbol{H}_0$, $\boldsymbol{H}_1$ *in* (3), *there exists an FIR-ZF equalizing filterbank* $\{\boldsymbol{G}_q\}_{q=0}^{Q-1}$ *so that:* $\boldsymbol{s}(n) = \sum_{q=0}^{Q-1} \boldsymbol{G}_q \boldsymbol{x}(n-q)$, *where the* $M \times P$ *matrix* $\boldsymbol{G}_q$ *is formed with columns* $\{\boldsymbol{g}_{q,p} = (g_p(qM) \cdots g_p(qM+M-1))^T\}_{p=0}^{P-1}$. *If in addition to* (**a0**)-(**a4**) *matrix* $\boldsymbol{G}_{Q-1} := (\boldsymbol{0}_{M\times L}|\tilde{\boldsymbol{G}}_{Q-1})$, *then the ZF equalizing filterbank* $\mathcal{G} := (\tilde{\boldsymbol{G}}_{Q-1}\, \boldsymbol{G}_{Q-2} \cdots \boldsymbol{G}_0)$ *is unique and is given by:*

$$\mathcal{G} = (\boldsymbol{0}_{M\times(Q-1)M}|\boldsymbol{I}_{M\times M})(\mathcal{H}(\boldsymbol{I}_{Q\times Q} \otimes \boldsymbol{F}))^{\dagger}, \tag{4}$$

where † *denotes pseudo-inverse, and* $\mathcal{H}$ *is a* $(QP-L) \times QP$ *Sylvester matrix with first row* $(h(L) \cdots h(0)0 \cdots 0)$. □

It is noteworthy that the existence and uniqueness conditions of Thm. 1 do not pose any constraints on the channel zeros. In contrast, FIR-ZF equalizers in [10], [11] and [9] do not exist for certain configurations of channel zeros on the unit circle, and more important, performance degrades even when channels have zeros close to those non-invertible configurations.

Given the transmission block size P and the channel order L, (a4) is met easily by selecting appropriately M and/or the matrix equalizer length Q. Minimum block size P requires $P = M+1$, and (a4) is then satisfied with equalizer length $Q \geq L$. On the other hand, simple zero-order ($Q = 1$) receiver filterbanks satisfy (a4) at the expense of extra redundancy: $P = M + L$ when M is fixed, or, with extra latency if both P and M have increase in order to maintain fixed information rate.

Assumption (a3) seems more technical but is satisfied by a number of special cases [7]. For spread-spectrum and CDMA systems, the $f_m(n)$ filters are often selected as pseudo-random and satisfy (a3) almost surely. TDMA systems have $\boldsymbol{F} = (\boldsymbol{I}_{M\times M}|\boldsymbol{0}_{M\times(P-M)})^T$ and will obey (a3), if $M \geq L$, because the M canonical vectors in $\boldsymbol{F}$ cannot be generated by less than M linearly independent vectors, and $M > L$. This corroborates the result in [3], where a TDMA precoder accepts a ZF equalizer without channel zero restrictions.

A more general class of precoders fulfilling (a3) results if one chooses the $P\times 1$ columns $\{f_m(n)\}_{m=0}^{M-1}$ of $\boldsymbol{F}$ from the M-dimensional subspace spanned by the canonical basis of $\mathcal{R}^P$. Because $P > L$, each of the canonical basis vectors is given as a linear combination of exactly P Vandermonde vectors, and hence (a3) is satisfied. Even when (a0)-(a3) hold and (a4) is satisfied as an equality, only existence (but not uniqueness) is guaranteed. Note that in Thm. 1 we also forced the equalizer filterbank to have leading zeros in order to assume uniqueness.

An alternative approach is to force the transmitter filterbank to have trailing zeros which also satisfies (a3) and leads to ZF equalization as we summarize next:

Theorem 2 (trailing precoder zeros): *Suppose that* (**a0**) *holds,* (**a1**) *is satisfied with* $P = M + L$, *and transmitter matrix* $\boldsymbol{F}$ *has trailing zeros (and thus* (**a3**) *holds), and also obeys* (**a2**)*; i.e.,* $\boldsymbol{F} := (\tilde{\boldsymbol{F}}^T_{M\times M}\ \boldsymbol{0}_{M\times L})^T$ *and* $rank(\tilde{\boldsymbol{F}}) = M$. *Then for a given* $\tilde{\boldsymbol{F}}$ *and channel matrix* $\boldsymbol{H}_0$, *there exists a zero-order* $(Q = 1)$ *ZF equalizer filterbank so that* $\boldsymbol{G}\boldsymbol{x}(n) = \boldsymbol{s}(n)$. *With* $\tilde{\boldsymbol{H}}_0$ *denoting the first* M *columns of* $\boldsymbol{H}_0$, *the minimum norm ZF filterbank is unique and is given by:*

$$\boldsymbol{G} = (\boldsymbol{H}_0\boldsymbol{F})^\dagger = \tilde{\boldsymbol{F}}^{-1}\tilde{\boldsymbol{H}}_0^\dagger \quad \square \tag{5}$$

Remark 1: Conventional OFDM does not satisfy (a3), because the columns of $\boldsymbol{F}$ are themselves Vandermonde vectors. However, Thm. 2 and the class of precoders with trailing zeros suggests a practical modification to the OFDM system. Instead of the cyclic prefix one could pad L trailing zeros to the M information bits. This allows for ZF equalizers even when channels have zeros on the unit circle located at angles $2\pi m/M$.

3 Transmitter-Receiver Optimal Designs

In the presence of noise, (3) becomes:

$$\boldsymbol{y}(n) = \boldsymbol{H}_0\boldsymbol{F}\boldsymbol{s}(n) + \boldsymbol{H}_1\boldsymbol{F}\boldsymbol{s}(n-1) + \boldsymbol{v}(n), \tag{6}$$

and the equalizers derived in the deterministic framework of Section 2 will be compared and augmented here with statistically optimal designs. Those will involve the signal and noise covariance matrices $\boldsymbol{R}_{ss}$ and $\boldsymbol{R}_{vv}$ which along with the channel matrices $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ will be assumed given. In contrast to Section 2, the precoder matrix $\boldsymbol{F}$ together with a zero order $(Q = 1)$ equalizer $\boldsymbol{G}$ will vary jointly to optimize our criteria of choice. As per Thm. 1, a zero-order equalizer (a4) requires $P = M + L$, which we assume herein.

3.1 Output SNR criterion: Our goal here is to maximize the SNR under the ZF constraint $\boldsymbol{G}\boldsymbol{H}_0\boldsymbol{F} = \mathcal{K}\boldsymbol{I}_{M\times M}$, where $\mathcal{K}$ is in general a complex constant whose magnitude $|\mathcal{K}| > 0$ depends on the transmitted power and controls the SNR at the equalizer output. Because the output SNR is only magnitude dependent, we take w.l.o.g. $\mathcal{K}$ real and positive. The same criterion was adopted in the scalar case (see e.g., [1]), but in our case the vector equalizer output is: $\hat{\boldsymbol{s}}(n) = \mathcal{K}\boldsymbol{s}(n) + \boldsymbol{G}\boldsymbol{v}(n)$, and the output noise variance becomes $tr(\boldsymbol{G}\boldsymbol{R}_{vv}\boldsymbol{G}^H)$. Our criterion maximizes SNR at every block sample n and is expressed as:

$$\max_{\boldsymbol{F},\boldsymbol{G}} \mathcal{K}^2 \frac{|tr(\boldsymbol{I}_{M\times M})|^2}{tr(\boldsymbol{G}\boldsymbol{R}_{vv}\boldsymbol{G}^H)} \quad \text{s. to} \quad \boldsymbol{G}\boldsymbol{H}_0\boldsymbol{F} = \mathcal{K}\boldsymbol{I}_{M\times M}. \tag{7}$$

Note that when $\boldsymbol{v}(n)$ is AGN, (7) is equivalent to minimizing the probability of error in a block-by-block detection scheme. The solution of (7) is summarized in the next theorem which is divided in two parts depending on whether the leading or trailing zeros are forced in the equalizer or in the precoder filterbank.

Theorem 3 (Max-SNR/ZF Equalizers): (**a**) *Let assumptions* (**a0**)-(**a4**) *be in force with* $P = M + L$ $(Q = 1)$ *and* $\boldsymbol{G} = (\boldsymbol{0}_{M\times L}\bar{\boldsymbol{G}}_{M\times M})$ *corresponding to an equalizer with leading zeros* (LZ). *Let also* $\bar{\boldsymbol{H}}_0$ $(\bar{\boldsymbol{v}}(n))$ *denote the last* M *rows of* $\boldsymbol{H}_0$ $(\boldsymbol{v}(n))$, *and* $\bar{\boldsymbol{R}}_{vv}$ *stand for the covariance matrix of* $\bar{\boldsymbol{v}}(n)$. *Define* $M \times M$ *diagonal matrix* $\bar{\boldsymbol{\Lambda}}$ *from the eigendecomposition* $\bar{\boldsymbol{H}}_0\bar{\boldsymbol{R}}_{vv}^{-1}\bar{\boldsymbol{H}}_0^H = \bar{\boldsymbol{V}}\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{V}}^H$. *The output SNR in* (7) *is then maximized by the optimum ZF pair of filterbanks:*

$$\boldsymbol{F}_{opt} = (\sqrt{\mathcal{K}}/\sigma_v)\, \bar{\boldsymbol{V}} \bar{\boldsymbol{\Lambda}}^{-\frac{1}{2}}, \qquad \bar{\boldsymbol{G}}_{opt} = \sigma_v \sqrt{\mathcal{K}}\, \bar{\boldsymbol{\Lambda}}^{-\frac{1}{2}} \bar{\boldsymbol{V}}^H \bar{\boldsymbol{H}}_0^H \bar{\boldsymbol{R}}_{vv}^{-1} . \tag{8}$$

(b) *Let assumptions* **(a0)**-**(a4)** *be in force with* $P = M + L$ $(Q = 1)$ *and* $\boldsymbol{F} = (\tilde{\boldsymbol{F}}_{M\times M}^T \boldsymbol{0}_{M\times L})^T$ *corresponding to a precoder with trailing zeros* (TZ)*. Define* $M \times M$ *diagonal matrix* $\tilde{\boldsymbol{\Lambda}}$ *from the eigendecomposition* $\tilde{\boldsymbol{H}}_0 \tilde{\boldsymbol{R}}_{vv}^{-1} \tilde{\boldsymbol{H}}_0^H = \tilde{\boldsymbol{V}} \tilde{\boldsymbol{\Lambda}} \tilde{\boldsymbol{V}}^H$ *and let* $\tilde{\boldsymbol{H}}_0$ *denote the first* M *columns of* $\boldsymbol{H}_0$*. The output SNR in* (7) *is then maximized by the optimum* TZ-ZF *pair of filterbanks:*

$$\tilde{\boldsymbol{F}}_{opt} = (\sqrt{\mathcal{K}}/\sigma_v) \tilde{\boldsymbol{V}} \tilde{\boldsymbol{\Lambda}}^{-\frac{1}{2}}, \quad \boldsymbol{G}_{opt} = \sigma_v \sqrt{\mathcal{K}}\, \tilde{\boldsymbol{\Lambda}}^{-\frac{1}{2}} \tilde{\boldsymbol{V}}^H \tilde{\boldsymbol{H}}_0^H \tilde{\boldsymbol{R}}_{vv}^{-1} . \quad \square \tag{9}$$

Interesting special cases appear when $\boldsymbol{v}(n)$ is white with variance σ_v^2. In this case, it suffices to replace $\bar{\boldsymbol{\Lambda}}$ with $\sigma_v^2 \bar{\boldsymbol{\Lambda}}$ in (8), to obtain

$$\boldsymbol{F}_{opt} = \sqrt{\mathcal{K}}\, \bar{\boldsymbol{V}} \bar{\boldsymbol{\Lambda}}^{-\frac{1}{2}}, \qquad \bar{\boldsymbol{G}}_{opt} = \sqrt{\mathcal{K}}\, \bar{\boldsymbol{U}}^H , \tag{10}$$

where $\bar{\boldsymbol{U}}$ is found from the SVD: $\bar{\boldsymbol{H}}_0 = \bar{\boldsymbol{U}} \bar{\boldsymbol{\Lambda}}^{\frac{1}{2}} \bar{\boldsymbol{V}}^H$. Similarly, when $\boldsymbol{v}(n)$ is white (9) becomes (after replacing $\tilde{\boldsymbol{\Lambda}}$ with $\sigma_v^2 \tilde{\boldsymbol{\Lambda}}$):

$$\tilde{\boldsymbol{F}}_{opt} = \sqrt{\mathcal{K}}\, \tilde{\boldsymbol{V}} \tilde{\boldsymbol{\Lambda}}^{-\frac{1}{2}}, \quad \boldsymbol{G}_{opt} = \sqrt{\mathcal{K}}\, \tilde{\boldsymbol{U}}^H , \tag{11}$$

where $\tilde{\boldsymbol{U}}$ is obtained from the SVD: $\tilde{\boldsymbol{H}}_0 = \tilde{\boldsymbol{U}} \tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}} \tilde{\boldsymbol{V}}^H$. In both (11) and (10) the optimum precoder (equalizer) filterbank matrices are proportional to the left (right) singular vectors of the channel matrix. Transmitter filterbanks are also weighted by the inverse of the square root of the corresponding singular values.

Remark 2: With flat fading (the channel matrix is diagonal) and white noise, (10) and (11) yield precoder and equalizer filterbanks with diagonal structure. Observing that TDMA corresponds to $f_m(n) = \delta(n - m)$, we infer that TDMA possesses optimality (in the ZF maximum output SNR sense) when it comes to channels involving flat fading and white noise.

3.2 Constrained Mean-Square Error Criterion: The ultimate criterion is to minimize error probability and although maximizing output SNR under the ZF constraint leads to simple closed form solutions, alternative criteria allowing residual ISI may come closer to the desired goal. One such candidate is the minimum mean-square error (MMSE) criterion which minimizes $\mathcal{E} := E\{tr(\boldsymbol{e}(n)\boldsymbol{e}^H(n))\}$, where $\boldsymbol{e}(n) := \hat{\boldsymbol{s}}(n) - \boldsymbol{s}(n)$ is the error in the nth data block. With $P = M + L$ and L trailing zeros, $\boldsymbol{y}(n) = \boldsymbol{H}_0 \boldsymbol{F} \boldsymbol{s}(n) + \boldsymbol{v}(n) = \tilde{\boldsymbol{H}}_0 \tilde{\boldsymbol{F}} \boldsymbol{s}(n) + \boldsymbol{v}(n)$, and hence:

$$\boldsymbol{e}(n) = \boldsymbol{G}\boldsymbol{y}(n) - \boldsymbol{s}(n) = \boldsymbol{G}\tilde{\boldsymbol{H}}_0 \tilde{\boldsymbol{F}} \boldsymbol{s}(n) + \boldsymbol{G}\boldsymbol{v}(n) - \boldsymbol{s}(n). \tag{12}$$

Using (12), the MSE objective function becomes:

$$\mathcal{E}(\boldsymbol{F}, \boldsymbol{G}) \;=\; tr((\boldsymbol{G}\tilde{\boldsymbol{H}}_0 \tilde{\boldsymbol{F}} - \boldsymbol{I})\boldsymbol{R}_{ss}(\boldsymbol{G}\tilde{\boldsymbol{H}}_0 \tilde{\boldsymbol{F}} - \boldsymbol{I})^H) + tr(\boldsymbol{G}\boldsymbol{R}_{vv}\boldsymbol{G}^H) . \tag{13}$$

But without any constraint, minimizing $\mathcal{E}$ leads to the trivial solution corresponding to $\|\boldsymbol{G}\| = 0$ and requiring infinite power to be transmitted $\|\boldsymbol{F}\| = \infty$. Imposing the ZF constraint not only leads to a nonlinear set of equations for $\boldsymbol{F}$ and $\boldsymbol{G}$, but also gives rise to infinite solutions indicating the need for extra restrictions in the available degrees of freedom. A reasonable alternative is to constrain the transmitted power expressed as $P_0 := tr(\tilde{\boldsymbol{F}} \boldsymbol{R}_{ss} \tilde{\boldsymbol{F}}^H)$. Our criterion thus becomes:

$$\min_{\tilde{\boldsymbol{F}}, \boldsymbol{G}} \mathcal{E}(\tilde{\boldsymbol{F}}, \boldsymbol{G}) \quad \text{s. to} \quad \mathcal{C} := tr(\tilde{\boldsymbol{F}} \boldsymbol{R}_{ss} \tilde{\boldsymbol{F}}^H) - P_0 = 0. \tag{14}$$

Analogous criteria formulated in the frequency domain for transmitter-receiver filter optimization can be found in the scalar case [1, p. 333], or, in the multi-input-multi-output case [12]. Optimizing our criterion in (14) follows the steps in [12], but our time-domain matrix formulation will lead to closed form selection of the FIR filterbank matrices (in [12], IIR frequency-domain designs are optimized via iterative minimization of Lagrange multipliers). Our result is summarized in the following theorem:

Theorem 4 (Constrained Power-MMSE Equalizer): *Let* (**a0**)-(**a4**) *hold true with* $P = M + L$ *and* $\boldsymbol{F} := (\tilde{\boldsymbol{F}}^T \quad \boldsymbol{0})^T$ *corresponding to a precoder with trailing zeros. Let also the channel matrix* $\boldsymbol{H}_0$ *be given and the diagonal matrices* $\boldsymbol{\Delta}$ $(\boldsymbol{\Lambda})$ *be determined by the eigen-decompositions:* $\boldsymbol{R}_{ss} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^H$, $\tilde{\boldsymbol{H}}_0^H \boldsymbol{R}_{vv}^{-1} \tilde{\boldsymbol{H}}_0 = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^H$, *where* $\boldsymbol{R}_{ss}$ $(\boldsymbol{R}_{vv})$ *denotes the information signal (noise) covariance matrix, which are also assumed to be available. Define the diagonal matrix* $\boldsymbol{\Phi}$ *with* (i,i) *entry:*

$$|\phi_i|^2 = \frac{P_0 + tr(\boldsymbol{\Lambda}^{-1})}{tr(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Delta}^{\frac{1}{2}})} \frac{1}{\sqrt{\lambda_i \delta_i}} - \frac{1}{\lambda_i \delta_i}, \tag{15}$$

where λ_i (δ_i) *is the* (i,i) *entry of* $\boldsymbol{\Lambda}$ $(\boldsymbol{\Delta})$. *The optimum* $(\boldsymbol{F}, \boldsymbol{G})$ *filterbank pair in the sense of (14) is given by:*

$$\tilde{\boldsymbol{F}}_{opt} = \boldsymbol{V}\boldsymbol{\Phi}\boldsymbol{U}^H, \qquad \boldsymbol{G}_{opt} = \boldsymbol{R}_{ss}\boldsymbol{F}_{opt}^H\boldsymbol{H}_0^H(\boldsymbol{R}_{vv} + \boldsymbol{H}_0\boldsymbol{F}_{opt}\boldsymbol{R}_{ss}\boldsymbol{F}_{opt}^H\boldsymbol{H}_0^H)^{-1} \tag{16}$$

and $\boldsymbol{F}_{opt} = (\tilde{\boldsymbol{F}}_{opt}^T \quad \boldsymbol{0})^T$ □.

Lower MMSE values may be reached only if one relaxes (**a2**) thereby allowing for certain symbols within the block to be non-recoverable from the received data.

4 Blind Symbol Recovery

Our discussion so far assumed that the ISI channel $\{h(l)\}_{l=0}^{L}$ is available. In certain applications (e.g., ADSL) the channel is estimated with input-output methods relying upon training sequences [5]. In wireless environments however, the multipath channel changes rapidly as the mobile communicators move, which motivates self-recovering equalization schemes to avoid frequent re-training and thus increase bandwidth efficiency. In this section, we derive blind channel estimation, direct equalization, and block synchronization schemes by exploiting the redundant filterbank precoders with TZs and $P = M + L$. As we mentioned earlier, the redundancy introduced by L zeros will have arbitrarily small reduction in the rate, $(P - M)/P$, if one chooses M sufficiently large. More important, as we saw in Thm. 2, such precoders lead to zero-order (and thus simple) equalizer filterbanks. We will seek first deterministic algorithms capable of reliable estimates even with short data records and later establish their consistency in the presence of noise. For symbol recovery we will assume (a0), but we will make (a1) and (a2) more specific:

(a1.1) For a given L, the pair (P, M) is chosen to satisfy **(a1)** and also $P \geq M + L$.

(a2.1) Precoders have L trailing zeros; i.e., $\boldsymbol{F} := (\tilde{\boldsymbol{F}}_{M \times M}^T \boldsymbol{0}_{M \times L})^T$ with $rank(\tilde{\boldsymbol{F}}) = M$.

We will also introduce a very mild (persistence-of-excitation) condition on the information sequence $s(n)$, namely that:

(a5) There exists an $N \geq P$, such that the $M \times N$ matrix $\boldsymbol{S} := (\boldsymbol{s}(0) \cdots \boldsymbol{s}(N-1))$, formed by the $\{\boldsymbol{s}(n)\}_{n=0}^{N-1}$ columns defined as in (3), has full rank M.

Note that with white inputs, $\boldsymbol{S}\boldsymbol{S}^H$ tends (as N increases) to the input correlation matrix $\boldsymbol{R}_{ss} = \sigma_v^2 \boldsymbol{I}$ which is full rank. But (a5) will be satisfied even for colored (e.g., coded) inputs provided that their spectra are non-zero for at least M frequencies (modes).

4.1 Blind Channel Estimation: Under (a1.1), eq. (3) becomes $\boldsymbol{x}(n) = \boldsymbol{H}_0\boldsymbol{F}\boldsymbol{s}(n) = \tilde{\boldsymbol{H}}_0\tilde{\boldsymbol{F}}\boldsymbol{s}(n)$, where $\tilde{\boldsymbol{H}}_0$ is formed by the first M columns of $\boldsymbol{H}_0$. Because $\tilde{\boldsymbol{H}}_0$ is a $P \times M$ lower triangular Toeplitz matrix with first column $(h(0)\cdots h(L)0\cdots 0)^T$, we will denote it as $\mathcal{T}(\boldsymbol{h})$ where $\boldsymbol{h} := (h(0)\cdots h(L))^T$. Collecting N data vectors $\{\boldsymbol{x}(n)\}_{n=0}^{N-1}$ in a matrix, we arrive at:

$$\boldsymbol{X}_N := (\boldsymbol{x}(0)\cdots\boldsymbol{x}(N-1)) = \tilde{\boldsymbol{H}}_0\,\tilde{\boldsymbol{F}}\,\boldsymbol{S}\,, \tag{17}$$

where $\boldsymbol{S}$ is defined as in (a5). Because of (a0), $h(0) \neq 0$, and thus $rank(\tilde{\boldsymbol{H}}_0) = M$, which along with (a2.1) and (a5) imply that $rank(\boldsymbol{X}_N) = M$. Therefore, the nullity of the matrix $\boldsymbol{X}_N\boldsymbol{X}_N^H$ is: $\nu(\boldsymbol{X}_N\boldsymbol{X}_N^H) = P - M = L$, and the eigen-decomposition

$$\boldsymbol{X}_N\boldsymbol{X}_N^H = \begin{pmatrix}\bar{\boldsymbol{U}} & \tilde{\boldsymbol{U}}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{M\times M} & \boldsymbol{0}_{M\times L}\\ \boldsymbol{0}_{L\times M} & \boldsymbol{0}_{L\times L}\end{pmatrix}\begin{pmatrix}\bar{\boldsymbol{U}}^H\\ \tilde{\boldsymbol{U}}^H\end{pmatrix}, \tag{18}$$

yields the $P \times L$ matrix $\tilde{\boldsymbol{U}}$ whose columns span the nullspace $\mathcal{N}(\boldsymbol{X}_N)$. Since $\tilde{\boldsymbol{F}}\boldsymbol{S}$ in (17) is full rank, $\mathcal{R}(\boldsymbol{X}_N) = \mathcal{R}(\boldsymbol{H}_0)$, where $\mathcal{R}$ stands for range space. But since $\mathcal{R}(\boldsymbol{X}_N)$ is orthogonal to $\mathcal{N}(\boldsymbol{X}_N)$, it follows that:

$$\tilde{\boldsymbol{U}}^H\,\tilde{\boldsymbol{H}}_0 = \boldsymbol{0} \;\Rightarrow\; \tilde{\boldsymbol{u}}_l^H\,\mathcal{T}(\boldsymbol{h}) = \boldsymbol{0}^H\,,\quad l = 1,\ldots,L \tag{19}$$

where $\tilde{\boldsymbol{u}}_l$ denotes the lth column of $\tilde{\boldsymbol{U}}$. But the vector multiplication with the Toeplitz matrix in (19) denotes convolution which is a commutative operation and thus (19) can be written as:

$$\boldsymbol{h}^H\,\boldsymbol{\mathcal{U}} := \boldsymbol{h}^H\,(\boldsymbol{\mathcal{U}}_1\cdots\boldsymbol{\mathcal{U}}_L) = \boldsymbol{0}^H\,, \tag{20}$$

where each $\boldsymbol{\mathcal{U}}_l$ is an $(L+1)\times M$ Hankel matrix formed by $\tilde{\boldsymbol{u}}_l$. From (20), the channel coefficient vector $\boldsymbol{h}$ can be obtained uniquely within a scale which is inherently non-identifiable by output only (blind) methods.

Our result and corresponding algorithm are summarized in the following:

Theorem 5 (Blind Channel Estimation): *Let* (a0), (a1.1), (a2.1) *and* (a5) *hold true. Starting from the data matrix* $\boldsymbol{X}_N$*, form the* $(L+1)\times ML$ *matrix* $\boldsymbol{\mathcal{U}}$ *as in (17)-(20). Channel vector* $\boldsymbol{h}$ *can then be obtained as the unique (within a scale) null eigen-vector of* $\boldsymbol{\mathcal{U}}$ *in (20).* □

With the channel matrix $\boldsymbol{H}_0$ available, we can proceed to determine either the ZF equalizer filterbank from (5), or, the MMSE equalizer from (16). In fact, it is possible to derive MMSE equalizers involving a delayed decision, or, pursue the computationally complex but optimal maximum-likelihood receiver. Instead, we focus next on direct blind equalizers that do not even require channel estimation as a first step and being linear, they lend themselves naturally to online self-recovering algorithms.

4.2 Direct Blind Equalization: Consider first the noise-free case and use the TZ condition (a2.1) to obtain

$$\boldsymbol{x}(n) = \boldsymbol{H}_0\boldsymbol{F}\boldsymbol{s}(n) = \boldsymbol{H}_0\begin{pmatrix}\tilde{\boldsymbol{F}}\boldsymbol{s}(n)\\ \boldsymbol{0}_{L\times 1}\end{pmatrix},\quad n = 0,1,\ldots,N-1\,. \tag{21}$$

Collect $\{\boldsymbol{x}(n)\}_{n=0}^{N-1}$ blocks to form the data matrix $\boldsymbol{X}_N$ as in (17), and define $\boldsymbol{\Gamma} := \boldsymbol{H}_0^{-1}$. Using (21), we then arrive at:

$$\boldsymbol{\Gamma}\,\boldsymbol{X}_N = \begin{pmatrix}\tilde{\boldsymbol{F}}\boldsymbol{S}(n)\\ \boldsymbol{0}_{L\times 1}\end{pmatrix}. \tag{22}$$

Because the $P \times P$ matrix $\boldsymbol{H}_0$ is lower triangular Toeplitz, it follows easily (by forming $\boldsymbol{\Gamma}\boldsymbol{H}_0 = \boldsymbol{I}$) that the inverse $\boldsymbol{\Gamma}$ is also lower triangular Toeplitz. Thus, all the rows $\{\gamma_i^T\}_{i=1}^{P}$

of $\boldsymbol{\Gamma}$ can be obtained from the last row $\boldsymbol{\gamma}_P^T$. Relying on (22), we will show how $\boldsymbol{\gamma}_P^T$ can be determined using only the received data matrix $\boldsymbol{X}_N$.

Let $\boldsymbol{J}$ denote a $P \times P$ shift matrix having all ones in the first sub-diagonal and all other entries zero. Using $\boldsymbol{J}$, the rows of $\boldsymbol{\Gamma}$ can be successively related: $\boldsymbol{\gamma}_i^T = \boldsymbol{\gamma}_{i+1}^T \boldsymbol{J}$ $\boldsymbol{\gamma}_{i+2}^T \boldsymbol{J}^2 = \cdots = \boldsymbol{\gamma}_P^T \boldsymbol{J}^{P-i}$, which implies that

$$\boldsymbol{\gamma}_P^T \boldsymbol{J}^{P-i} \boldsymbol{X}_N = \boldsymbol{\gamma}_i^T \boldsymbol{X}_N , \quad i = 1, \ldots, P . \tag{23}$$

But focusing on the last L rows of (22), it follows that for $i = P, P-1, \ldots, P-L+1$ we have $\boldsymbol{\gamma}_i^T \boldsymbol{X}_N = \boldsymbol{0}^T$, which after employing (23) leads to

$$\boldsymbol{\gamma}_P^T \boldsymbol{\mathcal{X}}_N := \boldsymbol{\gamma}_P^T \left(\boldsymbol{X}_N \, \boldsymbol{J} \boldsymbol{X}_N \cdots \boldsymbol{J}^{L-1} \boldsymbol{X}_N \right) = \boldsymbol{0}^T . \tag{24}$$

It can be shown that the nullity $\nu(\boldsymbol{\mathcal{X}}_N) = 1$; thus, $\boldsymbol{\gamma}_P^T$ (and hence $\boldsymbol{\Gamma}$) can be determined from (24) uniquely. With $\tilde{\boldsymbol{H}}_0$ denoting as before the first M columns of $\boldsymbol{H}_0$, it follows that the first M rows of $\boldsymbol{\Gamma}$ yield the pseudo-inverse $\boldsymbol{H}_0^\dagger$ which can be used in (5) to find the equalizer directly as:

$$\boldsymbol{G} = \tilde{\boldsymbol{F}}^{-1} \boldsymbol{H}_0^\dagger = \tilde{\boldsymbol{F}}^{-1} \left(\boldsymbol{I}_{M \times M} \, \boldsymbol{0}_{M \times M} \right) \boldsymbol{\Gamma} . \tag{25}$$

In summary, we maintain that:

Theorem 6 (Direct Blind Equalizer): *Let* **(a0)**, **(a1.1)**, **(a2.1)**, *and* **(a5)** *hold true. Then,* $\nu(\boldsymbol{\mathcal{X}}_N) = 1$, *and* $\boldsymbol{\gamma}_P^T$ *can be identified from (24) as the unique (within a scalar ambiguity) null eigen-vector of* $\boldsymbol{\mathcal{X}}_N \boldsymbol{\mathcal{X}}_N^H$. *With* $\boldsymbol{\gamma}_P^T$ *as the* P*th row, the lower triangular Toeplitz matrix* $\boldsymbol{\Gamma}$ *can be built and used in (25) to obtain the direct blind ZF equalizer* $\boldsymbol{G}$. □

The blind equalizer implied by Thm. 6 neither invokes any restrictions on the channel zeros nor it relies on any statistical input assumptions (e.g., whiteness as in [3]). It thus recovers the input exactly in the absence of noise. Because only $N \geq P$ data blocks are required in $\boldsymbol{X}_N$, and each $\boldsymbol{x}(n)$ is $P \times 1$, we infer that the minimum number of symbols required is P^2. Such a low requirement on the length of data records makes the direct equalizer a valuable self-recovering tool for rapidly fading multipath channels at least in high SNR environments.

4.3 Direct Blind Synchronization: So far we have assumed that block synchronization of the received data has been accomplished – a condition we adopted tacitly from the very beginning when we introduced our vector model in (3). Although techniques relying on training data can be deviced, it is possible to achieve block synchronization blindly – a task complementing our blind equalizer nicely. A statistical method exploiting the cyclostationarity of received samples, induced by the blocking operation at the transmitter, was presented in [9]. A deterministic blind synchronization algorithm is proposed herein after observing that matrix $\boldsymbol{\mathcal{X}}_N$ in (24) becomes full rank when the receiver is not block synchronous. It turns out that if blocks at the receiver are off by $d = t_0$ samples and one forms $\boldsymbol{\mathcal{X}}_N^{(d)}$ matrices for each possible shift d, then t_0 can be found as:

$$t_0 = \text{argmin}_d \, \lambda_{min}(\boldsymbol{\mathcal{X}}_N^{(d)} \boldsymbol{\mathcal{X}}_N^{(d)H}) , \tag{26}$$

where $\lambda_{min}(\boldsymbol{\mathcal{X}}_N^{(d)} \boldsymbol{\mathcal{X}}_N^{(d)H})$ is the minimum eigenvalue of $\boldsymbol{\mathcal{X}}_N^{(d)} \boldsymbol{\mathcal{X}}_N^{(d)H}$. In the noise-free case, we have $\lambda_{min}(\boldsymbol{\mathcal{X}}_N^{(t_0)} \boldsymbol{\mathcal{X}}_N^{(t_0)H}) = 0$ according to Thm. 6, and $\lambda_{min}(\boldsymbol{\mathcal{X}}_N^{(d)} \boldsymbol{\mathcal{X}}_N^{(d)H}) > 0$ for $d \neq t_0$.

4.4 Noisy Case – Consistency: When stationary additive noise $v(n)$ is present, our data model under (a2.1) is:

$$\boldsymbol{y}(n) = \boldsymbol{x}(n) + \boldsymbol{v}(n) = \tilde{\boldsymbol{H}}_0 \tilde{\boldsymbol{F}} \, \boldsymbol{s}(n) + \boldsymbol{v}(n) , \tag{27}$$

with corresponding $P \times P$ covariance matrix $\boldsymbol{R}_{yy} := E\{\boldsymbol{y}(n)\boldsymbol{y}^H(n)\}$ given by

$$\boldsymbol{R}_{yy} \;=\; \boldsymbol{R}_{xx} + \boldsymbol{R}_{vv} \;=\; \tilde{\boldsymbol{H}}_0 \tilde{\boldsymbol{F}} \boldsymbol{R}_{ss} \tilde{\boldsymbol{F}}^H \tilde{\boldsymbol{H}}_0 \;+\; \boldsymbol{R}_{vv} \,. \tag{28}$$

Similar to (a5), we assume that:

(a5.1) The $M \times M$ input covariance matrix is full rank; i.e., $rank(\boldsymbol{R}_{ss}) = M$.

Based on (a2.1) and (a5.1) we infer that $rank(\tilde{\boldsymbol{F}} \boldsymbol{R}_{ss} \tilde{\boldsymbol{F}}^H) = M$, whereas (a0) implies that $rank(\tilde{\boldsymbol{H}}_0) = M$. Hence, $rank(\boldsymbol{R}_{xx}) = M$, and $\nu(\boldsymbol{R}_{xx}) = P - M = L$, which yields an L-dimensional noise subspace $\tilde{\boldsymbol{U}}_x$ in the eigen-decomposition: $\boldsymbol{R}_{xx} = \boldsymbol{U}_x \boldsymbol{\Sigma}_x \boldsymbol{U}_x^H$, where $\boldsymbol{U}_x := (\bar{\boldsymbol{U}}_x \, \tilde{\boldsymbol{U}}_x)$. Arguing exactly as in (18), (19), we arrive at the orthogonality condition: $\tilde{\boldsymbol{U}}_x^H \tilde{\boldsymbol{H}}_0 = \boldsymbol{0}$. But since only noisy data are available, we proceed to relate $\tilde{\boldsymbol{U}}_x$ with $\tilde{\boldsymbol{U}}_y$, the 'noise subspace' of $\boldsymbol{R}_{yy}$. With this aim, let $\boldsymbol{R}_{vv} = \sigma_v^2 \boldsymbol{U}_v \boldsymbol{U}_v^H$ denote the eigen-decomposition of the noise covariance matrix which is assumed to be known. Based on $\boldsymbol{U}_v$, we eigen-decompose $\boldsymbol{U}_v^{-1} \boldsymbol{R}_{yy} \boldsymbol{U}_v^{-H}$ as:

$$\boldsymbol{U}_v^{-1} \boldsymbol{U}_x (\boldsymbol{\Sigma}_x + \sigma_v^2 \boldsymbol{I}) \boldsymbol{U}_v^H \boldsymbol{U}_v^{-H} := \boldsymbol{U}_y \boldsymbol{\Sigma}_y \boldsymbol{U}_y^{-1} \;= \left(\bar{\boldsymbol{U}}_y \, \tilde{\boldsymbol{U}}_y\right) \, \boldsymbol{\Sigma}_y \begin{pmatrix} \bar{\boldsymbol{U}}_y^H \\ \tilde{\boldsymbol{U}}_y^H \end{pmatrix} , \tag{29}$$

from which it follows easily that $\tilde{\boldsymbol{U}}_y = \boldsymbol{U}_v^{-1} \tilde{\boldsymbol{U}}_x$. Substituting the latter to our orthogonality condition, we find:

$$\tilde{\boldsymbol{U}}_y^H \, \boldsymbol{U}_v^H \, \tilde{\boldsymbol{H}}_0 \;=\; \boldsymbol{0} \,, \tag{30}$$

which implies that our blind channel identification algorithm applies even in the noisy case provided that $\boldsymbol{R}_{vv}$ (and hence $\boldsymbol{U}_v$) are known. Note also that knowledge of σ_v^2 is not necessary if $v(n)$ is white.

In practice, sample averages replace ensemble correlation matrices and consistency of our algorithm is guaranteed because $\boldsymbol{y}(n)$ in (27) is mixing. Indeed, input $s(n)$ has finite moments and the channel has finite memory; thus, relying upon standard law of large numbers arguments we have

$$\hat{\boldsymbol{R}}_{yy}^{(N)} \;:=\; \frac{1}{N} \sum_{n=0}^{N-1} \boldsymbol{y}(n) \boldsymbol{y}^H(n) \;\; \xrightarrow{\text{m.s.s.}} \;\; \boldsymbol{R}_{yy} \;, \;\; \text{as} \;\; N \to \infty \,, \tag{31}$$

where m.s.s. stands for mean-square convergence. Because $\tilde{\boldsymbol{U}}_y$ in (30) is continuous function of $\boldsymbol{R}_{yy}$, it follows that the estimate of $\tilde{\boldsymbol{U}}_y$ and thus the channel estimate will also be consistent.

The same argument establishes consistency of our deterministic approach in Section 4.1 when the noise $v(n)$ is white, if one observes that

$$\hat{\boldsymbol{R}}_{yy}^{(N)} \;=\; (1/N) \, \boldsymbol{Y}_N \boldsymbol{Y}_N^H \,, \tag{32}$$

where $\boldsymbol{Y}_N$ is formed exactly as $\boldsymbol{X}_N$ in (17) but with noisy data, and with subscript N signifying the number of blocks used.

Turning now back to the direct equalizer of Section 4.3, we find from (21) that:

$$\boldsymbol{\Gamma} \, \boldsymbol{R}_{xx} \;=\; \begin{pmatrix} \tilde{\boldsymbol{F}} \boldsymbol{R}_{ss} \tilde{\boldsymbol{F}}^H \tilde{\boldsymbol{H}}_0^H \\ \boldsymbol{0}_{L \times P} \end{pmatrix} \;\Rightarrow\; \bar{\gamma}_P^T \, \boldsymbol{R}_{xx} \;=\; \boldsymbol{0}^T \,, \tag{33}$$

and arguing as in (23) we infer that $\gamma_P^T \boldsymbol{J}^l \boldsymbol{R}_{xx} = \boldsymbol{0}^T$, which implies that $\gamma_P^T \boldsymbol{J}^l \boldsymbol{R}_{xx} (\boldsymbol{J}^l)^T = \boldsymbol{0}^T$, for $l = 0, 1, \ldots, L-1$. Hence,

$$\gamma_P^T \, \boldsymbol{\mathcal{R}}_{xx} \;=\; \boldsymbol{0}^T , \quad \boldsymbol{\mathcal{R}}_{xx} \;:=\; \sum_{l=0}^{L-1} \boldsymbol{J}^l \boldsymbol{R}_{xx} (\boldsymbol{J}^l)^T \,. \tag{34}$$

where $\mathcal{R}_{xx}$ can be shown to have nullity $\nu(\mathcal{R}_{xx}) = 1$. Thus, if noise were not present, γ_P (and hence $\boldsymbol{\Gamma}$) could be obtained (within a scale) as the null eigen-vector of $\mathcal{R}_{xx}$. In the noisy case, we define $\mathcal{R}_{yy}$ ($\mathcal{R}_{vv}$) as $\mathcal{R}_{xx}$ in (34) with $\boldsymbol{R}_{yy}$ ($\boldsymbol{R}_{vv}$) replacing $\boldsymbol{R}_{xx}$. Eigen-decomposing $\mathcal{R}_{vv} = \sigma_v^2 \mathcal{U}_v \mathcal{U}_v^H$, and using (28) we can write

$$\mathcal{U}_v^{-1} \mathcal{R}_{yy} \mathcal{U}_v^H \;=\; \mathcal{U}_v^{-1} \mathcal{R}_{xx} \mathcal{U}_v^H \;+\; \sigma_v^2 \boldsymbol{I}_{P\times P} \;=\; \mathcal{U}(\boldsymbol{\Lambda} + \sigma_v^2 \boldsymbol{I}) \mathcal{U}^H \;. \tag{35}$$

Next, we define $\bar{\gamma}_P^T := \gamma_P^T \mathcal{U}_v$, and use (34) to arrive at $\bar{\gamma}_P^T \mathcal{U}_v^{-1} \mathcal{R}_{xx} = \mathbf{0}^T$, which implies that $\bar{\gamma}_P^T \mathcal{U}_v^{-1} \mathcal{R}_{xx} \mathcal{U}_v^{-H} = \mathbf{0}^T$. The latter along with (35) yields

$$\bar{\gamma}_P^T \, \mathcal{U}_v^{-1} \mathcal{R}_{yy} \mathcal{U}_v^H \;=\; \sigma_v^2 \, \bar{\gamma}_P^T \,, \tag{36}$$

which shows that $\bar{\gamma}_P^T$ is the eigen-vector of $\mathcal{U}_v^{-1} \mathcal{R}_{xx} \mathcal{U}_v^{-H}$ corresponding to the minimum eigenvalue σ_v^2. Based on $\bar{\gamma}_P^T$ we can find $\gamma_P^T = \bar{\gamma}_P^T \mathcal{U}_v^{-1}$ and thus the direct equalizer $\boldsymbol{\Gamma}$ relying only on the noisy covariance matrix $\mathcal{R}_{yy}$ and knowledge of the noise covariance $\mathcal{R}_{vv}$. If $v(n)$ is white, then $\mathcal{R}_{vv} := \sigma_v^2 \sum_{l=0}^{L-1} \boldsymbol{J}^l (\boldsymbol{J}^l)^T \;=\; \sigma_v^2 \, diag(1\, 2 \cdots L \cdots L)$, and its spectral factor needed in (36) is given by: $\mathcal{U}_v \;:=\; diag(1\, \sqrt{2} \cdots L \cdots L)$.

With sample averages only available, consistency of our direct equalizer estimator follows from (31). The same argument establishes consistency of our deterministic method in Section 4.2 if we recall the definition of $\boldsymbol{Y}_N$ in (32) and note that

$$\hat{\mathcal{R}}_{yy}^{(N)} \;=\; (1/N)\, \mathcal{Y}_N \mathcal{Y}_N^H \,, \quad \mathcal{Y}_N \;:=\; \left(\boldsymbol{Y}_N \, \boldsymbol{J} \boldsymbol{Y}_N \cdots \boldsymbol{J}^{L-1} \boldsymbol{Y}_N \right) \,. \tag{37}$$

When the SNR is high, the deterministic solution should be preferred and the minimal number of blocks $N = P$ should be used for computational simplicity. However, at low SNR the statistical approach of this section will perform better provided that N is chosen large enough to obtain reliable estimates $\hat{\mathcal{R}}_{yy}^{(N)}$.

References

[1] S. Benedetto, E. Biglieri and V. Castellani, *Digital Transmission Theory,* Prentice-Hall Inc., New Jersey, 1987.

[2] L. J. Cimini, "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiple access," *IEEE Trans. on Communications*, pp. 665-675, 1995.

[3] G. B. Giannakis, "Filterbanks for blind channel identification and equalization," *IEEE Signal Processing Letters*, pp. 184-187, vol. 4, no. 6, June 1997.

[4] J. W. Lechleider, "The optimum combination of block codes and receivers for arbitrary channels," *IEEE Trans. on Communications*, pp. 615-621, May 1990.

[5] A. Ruiz, J.M. Cioffi, S. Kasturia, "Discrete multiple tone modulation with coset coding for the spectrally shaped channel", *IEEE Trans. on Comm.*, pp. 1012-1029, 1992.

[6] S. D. Sandberg , M.A.Tzannes, "Overlapped discrete multitone modulation for high speed copper wire communications" *IEEE J. on Sel. Areas in Com.*, pp. 1571-1585, 1995.

[7] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Redundant Filterbank Precoders and Equalizers, Part I: Unification and Optimal Designs; Part II: Blind Signal Recovery," *IEEE Trans. on Sig. Proc.*, 1997 (submitted).

[8] M. K. Tsatsanis, G. B. Giannakis, "Optimal linear receivers for DS-CDMA systems: a signal processing approach," *IEEE Trans. on Sig. Proc.*, pp. 3044-3055, 1996.

[9] M. K. Tsatsanis and G. B. Giannakis, "Transmitter Induced Cyclostationarity for Blind Channel Equalization," *IEEE Trans. on Sig. Proc.*, pp. 1785-1794, July 1997.

[10] L.Tong, G. Xu, T. Kailath, "Blind identification and equalization based on second order statistics: A time domain approach", *IEEE Trans. on IT*, pp.340-349, 1994.

[11] X.-G. Xia, "Applications of non-maximally decimated multirate filterbanks in partial response ISI cancellation," *Proc. of Intl. Symp. of TFTS,* Paris, France, 1996.

[12] J. Yang and S. Roy, "On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems," *IEEE Trans. on Comm.*, vol. 42, no. 12, pp. 3221-3231, December 1994.

On Infinite Families of 3-Designs from Preparata Codes and Related Codes over Z_4

Tor Helleseth [1], P. Vijay Kumar [2], Chunming Rong [3], and Kyeongcheol Yang [4]

Abstract

We show how to find several infinite families of 3-designs from the support of the codewords in the Preparata code and Kerdock code of length 2^m over Z_4 for all odd $m \geq 5$. Also the Goethals code over Z_4 contains examples of 3-designs for $m = 5$.

1 Introduction

A t-(v, k, λ) design D is a set of v points with a collection of k-subsets called blocks, such that any subset of t points is contained in precisely λ blocks. Many designs can be constructed from codes over a finite field F_q with q elements. The Assmus-Mattson theorem [1] gives necessary conditions for the support of the codewords of constant weight in a code to form a t-design.

For codes over Z_4 (the ring of integers modulo 4) no similar theorem has been found. Recently, Harada [3] was able to construct new 5-designs from the lifted Golay code over Z_4. These designs were constructed by a computer search.

The main purpose of this paper is to survey some infinite families of 3-designs obtained from Preparata codes, Kerdock code and Goethals codes over Z_4.

A linear code C over Z_4 with block length n is an additive subgroup of Z_4^n. The Lee weights of the elements $0, 1, 2, 3$ in Z_4 are $0, 1, 2, 1$, respectively. The Lee weight of a vector $\boldsymbol{a} \in Z_4^n$ is defined to be the sum of the Lee weights of its

---

[1] T. Helleseth is with the Department of Informatics, University of Bergen, Høyteknologisenteret, N-5020 Bergen, Norway.

[2] P. Vijay Kumar is with the Communication Sciences Institute, EE-Systems, University of Southern California, Los Angeles, CA 90089-2565, USA.

[3] C. Rong is with the Department of Informatics, University of Bergen, Høyteknologisenteret, N-5020 Bergen, Norway.

[4] K. Yang is with the Department of Electronic Communication Engineering, Hanyang University, Seoul 133-791, Korea.

components. The Gray map $\phi : Z_4 \to Z_2^2$ is defined by $\phi(0) = 00$, $\phi(1) = 01$, $\phi(2) = 11$, and $\phi(3) = 10$. In general, the binary code defined by $C = \phi(\mathcal{C})$ is a nonlinear binary code of length $2n$.

Let R_m be a Galois ring of characteristic 4 with 4^m elements and let R_m^* be the set of units of R_m. R_m^* has a multiplicative cyclic subgroup of order $2^m - 1$. Let $\mathcal{T}_m = \{0, 1, \beta, \ldots, \beta^{2^m-2}\}$, where $\beta \in R_m^*$ is an element of order $2^m - 1$. Any element $z \in R_m$ can be expressed uniquely as $z = A + 2B$ for $A, B \in \mathcal{T}_m$. Let μ denote the modulo-2 reduction map. Note that $\mu(\beta)$ is a primitive element in the finite field F_{2^m} with 2^m elements, thus $\mu(\mathcal{T}_m) = F_{2^m}$ (see [2] for details).

The Preparata code $\mathcal{P}_m$ of length 2^m over Z_4 is the code over Z_4, whose parity-check matrix is given by

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & \beta & \beta^2 & \cdots & \beta^{2^m-2} \end{bmatrix}.$$

In Hammons, Kumar, Calderbank, Sloane and Solé [2], it is shown that if m is odd, then $\mathcal{P}_m$ has minimum Lee weight 6 and its Gray map $P_m = \phi(\mathcal{P}_m)$ gives a $(2^{m+1}, 2^{2^{m+1}-2m-2}, 6)$ binary nonlinear code. The binary code P_m has the same Hamming weight distribution as the original Preparata code [7].

The support of a vector $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ is the subset of $\{1, 2, \ldots, n\}$ given by $\{i \mid a_i \neq 0\}$. If we select codewords of the same support size k from a code, it may be possible to construct t-designs with $v = n$ for an integer t.

We show that the support of codewords of support 5 and 6 in the Preparata code over Z_4 form a 3-designs for any odd integer $m \geq 3$. It is known in [8] (see also [7], p. 473) that the nonlinear binary Preparata code contains 3-designs. However, these 3-designs have different parameters than the 3-designs obtained from $\mathcal{P}_m$ in this paper.

2 Some Useful Lemmas

Let $(c_X)_{X\in\mathcal{T}_m}$ be a codeword of the Preparata code $\mathcal{P}_m$. Then it must satisfy

$$\sum_{X\in\mathcal{T}_m} c_X = 0$$

and

$$\sum_{X\in\mathcal{T}_m} c_X X = 0.$$

These relations give an invariant property of $\mathcal{P}_m$.

Lemma 1 ([2]) *The Preparata code defined in (1) is invariant under the doubly transitive group of "affine" permutations of the form*

$$X \to (AX + B)^{2^m}$$

where $A, B \in \mathcal{T}_m$ *and* $A \neq 0$.

The next lemma shows how an equation over R_m can be represented as two equivalent equations over F_{2^m}.

Lemma 2 ([4]) *Let* $\mathbf{e} = (e_X)_{X \in \mathcal{T}_m}$ *and let* $E_j = \{X \mid e_X = j\}$ *for* $j = 0, 1, 2, 3$. *The equation given by*

$$\sum_{X \in \mathcal{T}_m} e_X X = A + 2B, \quad A, B \in \mathcal{T}_m, e_X \in Z_4$$

is equivalent to the two binary equations

$$a = \sum_{X \in E_1 \cup E_3} x$$

$$b^2 = \sum_{X \in E_2 \cup E_3} x^2 + \sum_{\substack{X, Y \in E_1 \cup E_3 \\ X < Y}} xy$$

where $x, y, a,$ *and* b *are the modulo-2 reductions of* $X, Y, A,$ *and* B, *respectively, and* $<$ *is any ordering of the elements in* $\mathcal{T}_m$.

3 3-designs in Preparata code over Z_4

A vector is denoted to be of the type $1^{n_1}\, 2^{n_2}\, 3^{n_3}\, 0^{n_0}$ if i occurs n_i times, $i = 0, 1, 2, 3$, as a component. The codewords of minimum Lee weight in the Preparata code $\mathcal{P}_m$ for any odd integer m are of the type $1^3\, 2^1\, 3^1\, 0^{n-5}$ or $1^1\, 2^1\, 3^3\, 0^{n-5}$. Changing the sign of a codeword leads to a codeword with the same support. Hence, to construct simple designs (designs without repeated blocks), we only consider the former type. Note that the codewords of minimal Lee weight have Hamming weight 5.

Theorem 1 *The support of the codewords of the type* $1^3\,2^1\,3^1\,0^{n-5}$ *(or* $1^1\,2^1\,3^3\,0^{n-5}$*) in the Preparata code* $\mathcal{P}_m$ *over* Z_4 *form a* $3\text{-}(2^m, 5, 10)$ *design for any odd integer* $m \geq 3$.

Proof. We denote the codewords of the type $1^3 2^1 3^1 0^{n-5}$ as minimal and will show that for any three positions X_1, X_2 and X_3 in $\mathcal{T}_m$, there are exactly 10 minimal codewords with nonzero support in these positions. Since the Preparata code is invariant under the doubly-transitive group of "affine" permutations by Lemma 1, we can assume without loss of generality that $X_1 = 0$ and $X_2 = 1$. Furthermore, we let $X_3 = A$.

Let U_i denote the nonzero values of a minimal codeword at the five corresponding locations X_i for $i = 1, 2, \ldots, 5$. We will discuss the conditions for having a minimal codeword with this support. For a minimal codeword, there are 13 possible ways for the values U_1, U_2 and U_3 in the positions X_1, X_2, and X_3. The first seven cases are listed in Table I and the last six in Table II. We let $x_i = \mu(X_i)$ and $a = \mu(A)$.

TABLE I. FINDING POSITIONS x_4 AND x_5 FROM $x_1 = 0$, $x_2 = 1$ AND $x_3 = a$ FOR MINIMAL CODEWORDS IN THE PREPARATA CODE.

Case	$x_1 = 0$	$x_2 = 1$	$x_3 = a$	x_4	x_5	Comments
1	1	1	1	3	2	Unique solution
2	3	1	1	1	2	Unique solution
3	1	3	1	1	2	Unique solution
4	1	1	3	1	2	Unique solution
5	2	1	1	3	1	Unique solution
6	1	2	1	3	1	Unique solution
7	1	1	2	3	1	Unique solution

We first consider the cases in Table I. They can all easily be proved in a similar way and we discuss Case 1 to illustrate the method. Let $U_1 = 1$, $U_2 = 1$ and $U_3 = 1$. We will find a unique minimal codeword with $U_4 = 3$ and $U_5 = 2$ for some X_4 and X_5 in $\mathcal{T}_m$ that is uniquely determined.

The conditions lead to the following equation

$$X_1 + X_2 + X_3 + 3X_4 + 2X_5 = 0.$$

It is sufficient to determine $x_i \in F_{2^m}$. Since $x_1 = 0$, $x_2 = 1$, and $x_3 = a$, Lemma 2 shows that this is equivalent to the following equations over F_{2^m}:

$$1 + a + x_4 = 0,$$
$$a + x_4 + ax_4 + x_4^2 + x_5^2 = 0.$$

This leads to the unique solution $x_4 = a+1$ and $x_5 = \sqrt{a}$. Hence, there is a unique minimal codeword with the given support in X_1, X_2 and X_3. The Cases 2-7 are shown in the same way and the corresponding solutions are listed in Table I.

TABLE II. FINDING POSITIONS x_4 AND x_5 FROM $x_1 = 0$, $x_2 = 1$ AND $x_3 = a$ FOR MINIMAL CODEWORDS IN THE PREPARATA CODE.

Case	$x_1 = 0$	$x_2 = 1$	$x_3 = a$	x_4	x_5	Comments
8a	2	1	3	1	1	Exactly one solution
8b	2	3	1	1	1	in either 8a or 8b
9a	1	2	3	1	1	Exactly one solution
9b	3	2	1	1	1	in either 9a or 9b
10a	1	3	2	1	1	Exactly one solution
10b	3	1	2	1	1	in either 10a or 10b

In Table II the remaining six cases are considered. In these cases we will not always have a codeword with the given support in X_1, X_2 and X_3. However, it is possible to divide the six cases in three pairs $(8a, 8b)$, $(9a, 9b)$, $(10a, 10b)$, such that each pair has exactly one solution. For further details see [5]. The result follows since each of the 10 cases contains a unique minimal vector with the first three positions as its support. Hence, the number of minimal codewords with a support containing any three position is $\lambda = 10$. □

In [6] several other 3-designs can be found from codewords of support size 6 in the Preparata code $\mathcal{P}_m$ for any odd integer m. These codewords are of the type $1^5\, 2^0\, 3^1\, 0^{n-6}$ or $1^4\, 2^2\, 3^0\, 0^{n-6}$ or $1^3\, 2^0\, 3^3\, 0^{n-6}$ or $1^2\, 2^2\, 3^2\, 0^{n-6}$. From these codewords, we construct some simple designs (designs without repeated blocks).

Theorem 2 *The support of the codewords of type* $1^5\, 2^0\, 3^1\, 0^{n-6}$ *in the Preparata code* $\mathcal{P}_m$ *over* Z_4 *form a* 3-$(2^m,\ 6,\ 2^m - 8)$ *design for any odd integer* $m \geq 5$.

Theorem 3 *The support of the codewords of type* $1^4\, 2^2\, 3^0\, 0^{n-6}$ *in the Preparata code* $\mathcal{P}_m$ *over* Z_4 *form a* 3-$(2^m,\ 6,\ 5\cdot(2^{m-1}-4))$ *design for any odd integer* $m \geq 5$.

Theorem 4 *The support of the codewords of type* $1^3\,2^0\,3^3\,0^{n-6}$ *in the Preparata code* $\mathcal{P}_m$ *over* Z_4 *form a* 3-$(2^m,\,6,\,20\cdot(2^{m-1}-4)/3)$ *design for any odd integer* $m \geq 5$.

Theorem 5 *The support of the codewords of type* $1^2\,2^2\,3^2\,0^{n-6}$ *in the Preparata code* $\mathcal{P}_m$ *over* Z_4 *form a* 3-$(2^m,\,6,\,18\cdot(2^{m-1}-4))$ *design for any odd integer* $m \geq 5$.

4 Designs from Kerdock codes over Z_4

The Frobenius map σ from R_m to R_m is defined for any $z = a + 2b$ by $\sigma(z) = a^2 + 2b^2$ and the trace map from R_m to Z_4 is defined by

$$T(z) = \sum_{j=0}^{m-1} \sigma^j(z).$$

The Kerdock code $\mathcal{K}_m$ of length 2^m over Z_4 is defined by

$$\mathcal{K}_m = \{c(a,b) \mid a \in R_m,\ b \in Z_4\}$$

where $c(a,b)$ is a vector in $Z_4^{2^m}$ indexed by the elements of $\mathcal{T}_m$ such that $c(a,b)_x = T(ax) + b$ for all $x \in \mathcal{T}_m$. Clearly, $\mathcal{K}_m$ has 4^{m+1} codewords. In [2], Hammons, Kumar, Calderbank, Sloane, and Solé showed that if m is odd, then $\mathcal{K}_m$ has minimum Lee weight $2^m - 2^{\frac{m-1}{2}}$ and its Gray map $\phi(\mathcal{K}_m)$ gives a $(2^{m+1}, 2^{2m+2}, 2^m - 2^{\frac{m-1}{2}})$ binary nonlinear Kerdock code. The complete weight distribution of $\mathcal{K}_m$ is well-known in [2].

Lemma 3 *Let* m *be an odd integer* ≥ 3 *and let* $\mathcal{K}_m$ *be the Kerdock code of length* $n = 2^m$ *over* Z_4. *Then every codeword has one of the following types:*

(a) i^n, *one time for each* $i = 0, 1, 2, 3$;

(b) $2^{n/2}\,0^{n/2}$, $2(2^m-1)$ *times;*

(c) $1^{n/2}\,3^{n/2}$, $2(2^m-1)$ *times;*

(d) $1^{n_1}\,2^{n_2}\,3^{n_3}\,0^{n-n_1-n_2-n_3}$, $2^m(2^m-1)$ *times, where* $n_1 = 2^{m-2} \pm 2^{\frac{m-3}{2}}$, $n_2 = 2^{m-2} - 2^{\frac{m-3}{2}}$, *and* $n_3 = 2^{m-2} \mp 2^{\frac{m-3}{2}}$;

(e) $1^{n_1}\,2^{n_2}\,3^{n_3}\,0^{n-n_1-n_2-n_3}$, $2^m(2^m-1)$ *times, where* $n_1 = 2^{m-2} \pm 2^{\frac{m-3}{2}}$, $n_2 = 2^{m-2} + 2^{\frac{m-3}{2}}$, *and* $n_3 = 2^{m-2} \mp 2^{\frac{m-3}{2}}$.

Changing the sign of a codeword leads to a codeword with the same support. Hence, to construct simple designs (designs without repeated blocks), we only consider the codewords with $n_1 \geq n_3$ when $n_1 \neq 0$.

We show that the support of the codewords of each type in the Kerdock code $\mathcal{K}_m$ over Z_4 form 3-designs for any odd integer $m \geq 3$. Clearly, the codewords of the types given in Lemma 3 (a), (b) and (c) give trivial 3-designs. In [9] two new infinite families are constructed from the codewords of the types given in Lemma 3 (d) and (e), respectively.

It is well-known that the nonlinear binary Kerdock code contains 3-designs ([7], p. 456). However, these 3-designs have different parameters than the 3-designs obtained from the Kerdock code $\mathcal{K}_m$ over Z_4 in this paper.

Theorem 6 *The support of the codewords of the type* $1^{n_1}\,2^{n_2}\,3^{n_3}\,0^{n-n_1-n_2-n_3}$ *in the Kerdock code* $\mathcal{K}_m$ *of length* $n = 2^m$ *over* Z_4 *form a* 3-$(2^m, k, \lambda)$ *design for any odd integer* $m \geq 3$, *where*

$$k = 2^{m-1} + 2^{m-2} - 2^{\frac{m-3}{2}}, \quad \lambda = \frac{k(k-1)(k-2)}{2^m - 2}$$

and

$$n_1 = 2^{m-2} + 2^{\frac{m-3}{2}}, \quad n_2 = n_3 = 2^{m-2} - 2^{\frac{m-3}{2}}.$$

Theorem 7 *The support of the codewords of the type* $1^{n_1}\,2^{n_2}\,3^{n_3}\,0^{n-n_1-n_2-n_3}$ *in the Kerdock code* $\mathcal{K}_m$ *of length* $n = 2^m$ *over* Z_4 *form a* 3-$(2^m, k, \lambda)$ *design for any odd integer* $m \geq 3$, *where*

$$k = 2^{m-1} + 2^{m-2} + 2^{\frac{m-3}{2}}, \quad \lambda = \frac{k(k-1)(k-2)}{2^m - 2}$$

and

$$n_1 = n_2 = 2^{m-2} + 2^{\frac{m-3}{2}}, \quad n_3 = 2^{m-2} - 2^{\frac{m-3}{2}}.$$

The codewords in Theorem 6 are called minimal and the codewords in Theorem 7 are called maximal. We choose any three positions X_1, X_2 and X_3 in $\mathcal{T}_m$. We will count the number of minimal (or maximal) codewords with nonzero values in the positions X_1, X_2 and X_3. Since the Kerdock code is also known to be invariant under the doubly-transitive group of "affine" permutations given in Lemma 1, we can assume without loss of generality that $X_1 = 0$ and $X_2 = 1$. Furthermore, we let $X_3 = \tau$.

Let u_i denote the nonzero values of a minimal (or maximal) codeword $c(a, b)$ where $a = \gamma + 2\delta$ at the corresponding locations X_i for $i = 1, 2, 3$. Then we have

$$b = u_1,$$

$$b + T(\gamma + 2\delta) = u_2,$$

$$b + T((\gamma + 2\delta)\tau) = u_3,$$

and $b - T(\delta/\gamma) = s$ depending on m for minimality (or maximality) as shown in [10]. Therefore, we have the system of equations:

$$\begin{aligned} T(\gamma + 2\delta) &= u_2 - u_1, \\ T((\gamma + 2\delta)\tau) &= u_3 - u_1, \\ T\left(\frac{\delta}{\gamma}\right) &= u_1 - s \end{aligned} \tag{1}$$

where $u_1, u_2, u_3 \in Z_4^+$, $\gamma \in \mathcal{T}_m^*$, $\delta, \tau \in \mathcal{T}_m$, and $s \in Z_4$.

Let $N(u_1, u_2, u_3, \tau)$ be the number of solutions $a = \gamma + 2\delta$ to (1), given u_1, u_2, u_3 and τ. In order to construct 3-designs, it is necessary to show that

$$\sum_{u_1 \in Z_4^+} \sum_{u_2 \in Z_4^+} \sum_{u_3 \in Z_4^+} N(u_1, u_2, u_3, \tau)$$

is independent of τ, where $Z_4^+ = \{1, 2, 3\}$.

This can be shown by transferring this problem in a natural way to an exponential sum over R_m. Through a series of technical manipulations of this sum it follows that the above sum is independent of τ.

5 Conclusions

We have found several infinite families of 3-designs from the support of the codewords in the Preparata code and Kerdock code of length 2^m over Z_4 for all odd

$m \geq 5$. Preliminary results also reveal examples of 3-designs in the Goethals code over Z_4 (defined in [2]) for $m = 5$.

References

[1] E. F. Assmus and H. F. Mattson, Jr., New 5-designs, *J. Combin. Theory. Theory*, Vol. 6 (1969) pp. 122-151.

[2] A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Sole, The Z_4-linearity of Kerdock, Preparata, Goethals, and related codes, *IEEE Trans. Inform. Theory*, Vol. 40 (1994) pp. 301-319.

[3] M. Harada, New 5-designs constructed from the lifted Golay code over Z_4, preprint, presented at the Recent Results Session of *1997 IEEE International Symposium on Information Theory*, Ulm, Germany, June 29-July 4, 1997.

[4] T. Helleseth and P. V. Kumar, The algebraic decoding of the Z_4-linear Goethals code, *IEEE Trans. Inform. Theory*, Vol. 41 (1995) pp. 2040-2048.

[5] T. Helleseth, P. V. Kumar and K. Yang, An infinite family of 3-designs from Preparata codes over Z_4, submitted for publication.

[6] T. Helleseth, C. Rong and K. Yang, 3-designs from Preparata codes over Z_4, Report in Informatics, University of Bergen, No. 137 (1997).

[7] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland (1977).

[8] N. V. Semakov, V. A. Zinoviev and G. V. Zaitsev, Uniformly packed codes, *Problems of Inform. Transmission*, Vol. 7 (1971) pp. 30-39.

[9] K. Yang and T. Helleseth, Two new families of 3-designs from Kerdock codes over Z_4, submitted for publication.

[10] K. Yang and T. Helleseth, On the weight hierarchy of Kerdock codes over Z_4, *IEEE Trans. Inform. Theory*, Vol. 42 (1996) pp. 1587-1594.

The Decomposition of $\mathbb{Z}_4$ Codes

W. Cary Huffman
Department of Mathematical and Computer Sciences
Loyola University
Chicago, Illinois

Abstract

If a linear $\mathbb{Z}_4$-code has an automorphism of odd order, the code can be decomposed as a direct sum of subcodes which possess useful properties for studying the original code. In particular, the subcodes can be viewed as shorter length codes over Galois rings extending $\mathbb{Z}_4$. If in addition, the original code is self-dual, the component codes either 'almost' self-dual or 'almost' dual in pairs. Using the decomposition of $\mathbb{Z}_4$-codes, it is sometimes quite easy to decide on equivalence or inequivalence of two codes. We apply this theory to the study of self-dual $\mathbb{Z}_4$-codes of length 24 with minimum Euclidean weight 16. There are 42 such codes with automorphisms of prime order 5 or more. These lead to 42 constructions of the Leech lattice.

1 Introduction

In this paper we extend the code decomposition theory developed in [7, 8, 9, 13] for self-dual linear codes over $F_q = \mathrm{GF}(q)$ to self-dual codes over $\mathbb{Z}_4$, the ring of integers modulo 4. We will then apply the theory to Type II codes of length 24. A complete description of the results can be found in [10].

Before presenting the decompositions, we need a brief discussion of the Galois ring $\mathrm{GR}(4^m)$, which is the unique Galois extension of degree m of $\mathbb{Z}_4$. (General results on Galois rings can be found in [2, 6, 11, 12].) One method of constructing $\mathrm{GR}(4^m)$ is as follows. Let X be an indeterminate. A polynomial $f(X) \in \mathbb{Z}_4[X]$ is *basic irreducible* if $f(X) \pmod 2$ is irreducible in $F_2[X]$ and is *monic* if its leading coefficient is 1. If $f(X)$ is any monic basic irreducible polynomial of degree m, then $\mathbb{Z}_4[X]/(f(X)) \simeq \mathrm{GR}(4^m)$ where $(f(X))$ is the principal ideal of $\mathbb{Z}_4[X]$ generated by $f(X)$. $\mathcal{R} = \mathrm{GR}(4^m)$ contains an element ξ of order $2^m - 1$, called a *primitive element.* Every element $c \in \mathcal{R}$ can be uniquely expressed in its 2-*adic* form $c = a + 2b$ where a and b are elements of

$$\mathcal{T}(\mathcal{R}) = \{0, 1, \xi, \xi^2, \ldots, \xi^{2^m-2}\}.$$

$\mathcal{R}$ has only three ideals, namely $(0) \subset (2) = 2\mathcal{R} = \{2t \mid t \in \mathcal{T}(\mathcal{R})\} \subset (1) = \mathcal{R}$. The elements of $\mathcal{R} \setminus (2)$ are precisely the invertible elements of $\mathcal{R}$, and the remaining nonzero elements are the divisors of 0 in $\mathcal{R}$. Also $\mathcal{R}/(2) \simeq \mathrm{GF}(2^m)$ with the coset $\xi + (2)$ corresponding to a primitive element of $\mathrm{GF}(2^m)$. The Galois group of $\mathrm{GR}(4^m)$ over $\mathbb{Z}_4$ is cyclic of order m generated by the *Frobenius map* ν_2 defined by

$$\nu_2(c) = a^2 + 2b^2,$$

where $a + 2b$ is the 2-adic representation of c.

Our code decomposition is based on the decomposition of the ring $\mathfrak{R}_r = \mathbb{Z}_4[X]/(X^r - 1)$ where r is an odd integer. Lemma 1.1 shows that $\mathfrak{R}_r$ is a direct sum of Galois rings and describes certain isomorphisms and automorphisms of these components. In what follows, when exhibiting principal ideals, (f) will denote the ideal generated by f in $\mathbb{Z}_4[X]$, while $\langle f\rangle$ will denote the ideal generated by f in $\mathfrak{R}_r$. The polynomial X^r-1 can be factored in $\mathbb{Z}_4[X]$ (uniquely up to order and multiplication by units) as a product

$$X^r-1=m_0(X)m_1(X)\cdots m_g(X)$$

where $m_i(X)$ are monic basic irreducible polynomials for $0\le i\le g$ with $m_0(X)=X-1$. Modulo 2, this is the factorization of X^r-1 into irreducible polynomials over F_2. Let $\widehat{m}_j(X)=(X^r-1)/m_j(X)$, and $\mathcal{I}_j=\langle\widehat{m}_j(X)\rangle$ for $0\le j\le g$. Notice that $\mathcal{I}_0=\{a(1+X+\cdots+X^{r-1})\mid a\in\mathbb{Z}_4\}\simeq\mathbb{Z}_4$ under the map $a(1+X+\cdots+X^{r-1})\to ar$. For any positive integer u relatively prime to r, define the map $\tau_u:\mathfrak{R}_r\to\mathfrak{R}_r$ by

$$\tau_u\left(\sum_{i=0}^{r-1}a_iX^i\right)=\sum_{i=0}^{r-1}a_iX^{ui}. \tag{1}$$

Since $(X^r-1)\mid(X^{ru}-1)$ and $\gcd(r,u)=1$, τ_u is a well-defined ring automorphism. Since in $\mathfrak{R}_r$ we can read exponents modulo r, the definition of τ_u extends to negative values of u relatively prime to r. The following is proved in [10].

Lemma 1.1 *The following hold:*

(i) $\mathfrak{R}_r=\mathcal{I}_0\oplus\mathcal{I}_1\oplus\cdots\oplus\mathcal{I}_g$.

(ii) $\langle X-1\rangle=\mathcal{I}_1\oplus\mathcal{I}_2\oplus\cdots\oplus\mathcal{I}_g$

(iii) *For* $0\le j\le g$, $\mathcal{I}_j=\langle\widehat{m}_j(X)\rangle\simeq\mathbb{Z}_4[X]/(m_j(X))\simeq GR(4^{d_j})$ *where* d_j *is the degree of* $m_j(X)$.

(iv) τ_u *is the identity on* $\mathcal{I}_0$.

(v) τ_u *permutes* $\mathcal{I}_1,\ldots,\mathcal{I}_g$ *and if* $\tau_u(\mathcal{I}_i)=\mathcal{I}_j$, *then* τ_u *is a ring isomorphism from* $\mathcal{I}_i$ *onto* $\mathcal{I}_j$.

The next two examples illustrate this decomposition of $\mathfrak{R}_r$. The polynomial $a_0+a_1X+\cdots+a_{r-1}X^{r-1}$ in $\mathfrak{R}_r$ is denoted $a_0a_1\cdots a_{r-1}$.

Example 1.2 Let $r=3$. Then $X^3-1=(3+X)(1+X+X^2)$, with each factor a basic irreducible. The ideal $\mathcal{I}_0=\langle 1+X+X^2\rangle=\{000,333,222,111\}\simeq\mathbb{Z}_4$ where the identity is $e_0=3+3X+3X^2$. The ideal $\mathcal{I}_1=\langle 3+X\rangle\simeq\mathrm{GR}(4^2)$. A primitive element is $\xi=1+2X+X^2$ and the identity is $\xi^3=\xi^0=2+X+X^2$. The divisors of zero are $2\xi^0=022$, $2\xi=202$, and $2\xi^2=220$. The 12 invertible elements are ξ^i and $\xi^i+2\xi^j$ where $0\le i\le 2$, $0\le j\le 2$; these are the polynomials $a_0+a_1X+a_2X^2$ with some $a_i\in\{1,3\}$. The element τ_2 maps ξ to ξ^2 and generates the Galois group of $\mathcal{I}_1$, which has order 2. Notice that $X=010=333+121=e_0+\xi$. Therefore multiplying an element of $\mathcal{I}_1$ by ξ is equivalent to a cyclic shift to the right one position.

Example 1.3 Now let $r=7$. Then

$$X^7-1=(3+X)(3+2X+3X^2+X^3)(3+X+2X^2+X^3)$$

is the factorization of X^7-1 into monic basic irreducibles. $\mathcal{I}_0=\langle 1+X+\cdots+X^6\rangle\simeq\mathbb{Z}_4$ has identity $e_0=3+3X+\cdots+3X^6$. The ideals $\mathcal{I}_1$ and $\mathcal{I}_2$ are both isomorphic to

GR(4^3). $\mathcal{I}_i$ has a 7th root of unity ξ_i such that $\mathcal{I}_i = \{a + 2b \mid a, b \in \{0, \xi_i^0, \ldots, \xi_i^6\}\}$. The following table gives enough information to construct the ideals.

i	$\widehat{m}_i$	ξ_i	identity $= \xi_i^0$
1	1231100	3122323	1223233
2	1132100	3323221	1332322

The map $\tau_6 = \tau_{-1}$ interchanges $\mathcal{I}_1$ and $\mathcal{I}_2$ interchanging ξ_1 and ξ_2. The maps τ_1, τ_2, and $\tau_4 = \tau_2^2$ make up the Galois group of both $\mathcal{I}_1$ and $\mathcal{I}_2$. Also $X = 0100000 = 3333333 + 3122323 + 2133232 = e_0 + \xi_1 + \xi_2^6$. So multiplying an element in $\mathcal{I}_1$ by ξ_1 and an element in $\mathcal{I}_2$ by ξ_2^6 cyclically shifts the element one to the right. Therefore ξ_1^j is obtained by cyclically shifting ξ_1^0 j places to the right; ξ_2^j is obtained by cyclically shifting ξ_2^0 j places to the left.

2 Code Decompositions

We are now ready to present the code decomposition. It comes directly from the decomposition in Lemma 1.1. Theorem 2.1 will show that a linear $\mathbb{Z}_4$-code possessing a permutation automorphism of odd prime order is the direct sum of subcodes, which can be viewed as shorter codes over Galois rings.

We need some terminology. First let $\mathcal{M}_n$ be the group of all invertible $n \times n$ monomial matrices with entries in $\mathbb{Z}_4$. If $M \in \mathcal{M}_n$, then $M = PD$ where P is an $n \times n$ permutation matrix, called the *permutation part* of M, and $D = \mathrm{diag}(a_1, a_2, \ldots, a_n)$ is an $n \times n$ diagonal matrix, called the *diagonal part* of M, with diagonal entries $a_i \in \{\pm 1\}$. Two linear $\mathbb{Z}_4$-codes $\mathcal{C}_1$ and $\mathcal{C}_2$ of length n are *equivalent* provided $\mathcal{C}_1 = \mathcal{C}_2 M$ for some $M \in \mathcal{M}_n$. Additionally, $\mathrm{Aut}(\mathcal{C}) = \{M \in \mathcal{M}_n \mid \mathcal{C}M = \mathcal{C}\}$ is the *automorphism group* of the $\mathbb{Z}_4$-code $\mathcal{C}$.

Suppose that $\mathcal{C}$ is a linear $\mathbb{Z}_4$-code of length n which has an automorphism σ of odd prime order r. By a simple replacement of $\mathcal{C}$ by an equivalent code, one can assume without loss of generality that σ has trivial diagonal part and is a permutation with c r-cycles and $f = n - cr$ fixed points where

$$\sigma = (1, 2, \ldots, r) \cdots ((c-1)r + 1, (c-1)r + 2, \ldots, cr). \tag{2}$$

(Later we will assume that $\mathcal{C}$ is self-dual, and it will be clear that any code equivalent to a self-dual code is also self-dual.) Denote the orbits of σ by $\Omega_i = \{(i-1)r + 1, (i-1)r + 2, \ldots, ir\}$ for $1 \leq i \leq c$ and $\Omega_{c+j} = \{cr + j\}$ for $1 \leq j \leq f$. Let $v|_{\Omega_i}$ denote the restriction of $v \in \mathbb{Z}_4^n$ to Ω_i. For $1 \leq i \leq c$, $v|_{\Omega_i}$ can be viewed as an element $a_0 + a_1 X + \cdots + a_{r-1} X^{r-1} \in \mathfrak{R}_r$, where $v\sigma|_{\Omega_i} = (a_0 + a_1 X + \cdots + a_{r-1} X^{r-1}) X \in \mathfrak{R}_r$. Define $\mathcal{C}(\sigma) = \{v \in \mathcal{C} \mid v\sigma = v\}$. For $1 \leq i \leq g$, let $\mathcal{E}_i(\sigma) = \{v \in \mathcal{C} \mid v|_{\Omega_j} \in \mathcal{I}_i$ for $1 \leq j \leq c$ and $v|_{\Omega_j} = 0$ for $c + 1 \leq j \leq c + f\}$. The following theorem gives our code decomposition; its proof is in [10].

Theorem 2.1 *Let $\mathcal{C}$ be a linear $\mathbb{Z}_4$-code with automorphism σ of odd prime order r as in* (2). *Then $\mathcal{C} = \mathcal{C}(\sigma) \oplus \mathcal{E}_1(\sigma) \oplus \cdots \oplus \mathcal{E}_g(\sigma)$.*

Each $\mathcal{E}_i(\sigma)$ can be viewed as a code over the Galois ring $\mathcal{I}_i$. Therefore, let $\mathcal{E}_i(\sigma)^*$ denote the code formed from $\mathcal{E}_i(\sigma)$ by puncturing on the fixed points $\Omega_{c+1} \cup \cdots \cup \Omega_{c+f}$ and considering the codewords as c-tuples with components in $\mathcal{I}_i$. Each $\mathcal{E}_i(\sigma)^*$ is 'linear' over $\mathcal{I}_i$ (that, it is an $\mathcal{I}_i$-module) as applying σ is the same as multiplication by X and

$\mathcal{I}_i$ is an ideal of $\mathfrak{R}_r$. In addition, the entries of $\mathcal{C}(\sigma)$ are constant on each r-cycle and so these entries correspond to elements of $\mathcal{I}_0$. Therefore define the $\mathbb{Z}_4$-module isomorphism $\pi : \mathcal{C}(\sigma) \to \mathbb{Z}_4^{c+f}$ by the natural projection

$$\pi(a_1 \sum_{i=0}^{r-1} X^i, \ldots, a_c \sum_{i=0}^{r-1} X^i, a_{c+1}, \ldots, a_{c+f}) = (a_1, \ldots, a_c, a_{c+1}, \ldots, a_{c+f}).$$

Let $\mathcal{C}(\sigma)^*$ denote $\pi(\mathcal{C}(\sigma))$.

The converse of Theorem 2.1 holds in the following sense. Namely if $\mathcal{C}$ can be decomposed as in the theorem and each $\mathcal{E}_i(\sigma)^*$ is an $\mathcal{I}_i$-module, then σ is indeed an automorphism of $\mathcal{C}$.

Example 2.2 Let $\mathcal{C} = \mathcal{O}_8$ be the octacode, the code whose Gray map image gives the Nordstrom–Robinson code [6]. This code has automorphism group of order 1344 with two generator matrices given in [4]. By examining the second generator matrix, an obvious automorphism of order $r = 3$ is found. By rearranging the order of coordinates and row reducing, we obtain the generator matrix with permutation automorphism $\sigma_3 = (1,2,3)(4,5,6)$:

$$G_3 = \left[\begin{array}{ccc|ccc|c|c} 1 & 1 & 1 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 \\ \hline 2 & 1 & 1 & 0 & 3 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 & 0 & 3 & 0 & 0 \end{array}\right].$$

The generator matrix for $\mathcal{C}(\sigma_3)^*$ is

$$\text{gen}(\mathcal{C}(\sigma_3)^*) = \left[\begin{array}{cccc} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{array}\right].$$

The generator matrix for $\mathcal{E}_1(\sigma_3)^*$ is

$$\text{gen}(\mathcal{E}_1(\sigma_3)^*) = \left[\begin{array}{cc} \xi^0 & \xi^0 + 2\xi^2 \end{array}\right]$$

because $\xi^0 = 211 = 2 + X + X^2$ and $\xi^0 + 2\xi^2 = 031 = 3X + X^2$ by Example 1.2. Notice that the last two rows of G_3 generate the code $\mathcal{E}_1(\sigma_3)$ of order 4^2. The single row of $\text{gen}(\mathcal{E}_1(\sigma_3)^*)$ also generates a code of order 4^2 as $\mathcal{E}_1(\sigma_3)^*$ is a code over $\mathcal{I}_1 \simeq \text{GR}(4^2)$.

The first generator matrix in [4] gives $\mathcal{O}_8$ as an extended cyclic code. By row reducing we obtain the generator matrix with permutation automorphism $\sigma_7 = (1,2,3,4,5,6,7)$:

$$G_7 = \left[\begin{array}{ccccccc|c} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 3 & 3 & 2 & 3 & 2 & 2 & 0 \\ 2 & 1 & 3 & 3 & 2 & 3 & 2 & 0 \\ 2 & 2 & 1 & 3 & 3 & 2 & 3 & 0 \end{array}\right].$$

The generator matrix for $\mathcal{C}(\sigma_7)^*$ is

$$\text{gen}(\mathcal{C}(\sigma_7)^*) = \left[\begin{array}{cc} 1 & 1 \end{array}\right].$$

The code $\mathcal{E}_1(\sigma_7)^*$ is $\{0\}$ and the generator matrix for $\mathcal{E}_2(\sigma_7)^*$ is

$$\text{gen}(\mathcal{E}_2(\sigma_7)^*) = \left[\begin{array}{c} \xi_2^0 \end{array}\right]$$

because $\xi_2^0 = 1332322$ by Example 1.3. Again the last three rows of G_7 generate the code $\mathcal{E}_2(\sigma_7)$ of order 4^3, which is the order of the code $\mathcal{E}_2(\sigma_7)^*$ over $\mathcal{I}_2 \simeq \text{GR}(4^3)$.

We now examine the consequences of Theorem 2.1 when applied to self-dual $\mathbb{Z}_4$-codes (which automatically must be linear). Let $\mathcal{S}$ be a ring. Define the *ordinary inner product* $\langle u, v\rangle_{\mathcal{S}}$ of vectors $u = (u_1, \ldots, u_m)$ and $v = (v_1, \ldots, v_m)$ in $\mathcal{S}^m$ by

$$\langle u, v\rangle_{\mathcal{S}} = \sum_{i=1}^{m} u_i v_i.$$

The *dual* $\mathcal{C}^\perp$ of an $\mathcal{S}$-submodule $\mathcal{C}$ of $\mathcal{S}^m$ (that is, a 'linear' $\mathcal{S}$-code of length m) is defined by

$$\mathcal{C}^\perp = \{u \in \mathcal{S}^m \mid \langle u, v\rangle_{\mathcal{S}} = 0 \text{ for all } v \in \mathcal{C}\}.$$

As usual, $\mathcal{C}$ is *self-orthogonal* if $\mathcal{C} \subseteq \mathcal{C}^\perp$ and *self-dual* if $\mathcal{C} = \mathcal{C}^\perp$.

By Lemma 1.1(v), τ_{-1} induces a permutation λ of $1, 2, \ldots, g$ given by

$$\tau_{-1}(\mathcal{I}_i) = \mathcal{I}_{\lambda(i)}. \tag{3}$$

Notice that λ consists a product of 1- or 2-cycles. For instance, when $r = 3$, λ is clearly (1) by Example 1.2; when $r = 7$, τ_{-1} swaps $\mathcal{I}_1$ and $\mathcal{I}_2$ giving $\lambda = (1, 2)$ from Example 1.3. The following is proved in [10].

Theorem 2.3 *Assume $\mathcal{C}$ is a self-dual $\mathbb{Z}_4$-code with automorphism σ as in* (2). *Then $\mathcal{C}(\sigma)$ is self-orthogonal with 2^{c+f} codewords and $\mathcal{E}_{\lambda(i)}(\sigma)^* = (\tau_{-1}(\mathcal{E}_i(\sigma)^*))^\perp$ under $\langle \cdot, \cdot\rangle_{\mathcal{I}_{\lambda(i)}}$ for all i with $1 \le i \le g$.*

So when $\mathcal{C}$ is self-dual, the codes $\mathcal{E}_i(\sigma)^*$ are either 'almost' self-dual or 'almost' dual in pairs. In other words, $\mathcal{E}_i(\sigma)^*$ is either dual to the τ_{-1} image of itself or another $\mathcal{E}_j(\sigma)^*$.

Example 2.4 Returning to Example 2.2, the octacode is self-dual. When $r = 3$, $\mathcal{C}(\sigma_3)$ is self-orthogonal with 2^4 codewords. Also $\mathcal{E}_1(\sigma_3)^*$ and $\tau_{-1}(\mathcal{E}_1(\sigma_3)^*)$ are dual because if $v = (\xi^0, \xi^0+2\xi^2)$, then $\tau_{-1}(v) = (\xi^0, \xi^0+2\xi^1)$ and therefore $\langle v, \tau_{-1}(v)\rangle_{\mathcal{I}_1} = 2\xi^0+2\xi^1+2\xi^2 = 0$. When $r = 7$, $\mathcal{C}(\sigma_7)$ is self-orthogonal with 2^2 codewords, and $\mathcal{E}_2(\sigma_7)^*$ and $\tau_{-1}(\mathcal{E}_1(\sigma_7)^*)$ are clearly duals.

We remark that if $r \equiv 1 \pmod 4$ and $\mathcal{C}$ is self-dual, then an easy consequence of Theorem 2.3 is that $\mathcal{C}(\sigma)^*$ is a self-dual $\mathbb{Z}_4$-code. Also the converse of Theorem 2.3 holds.

We turn to the question of when two codes are equivalent. Suppose there are two linear $\mathbb{Z}_4$-codes $\mathcal{C}$ and $\mathcal{C}'$ each having the same automorphism σ of prime order r and form (2). Then, under the assumption that $\langle\sigma\rangle$ is a Sylow r-subgroup of $\mathrm{Aut}(\mathcal{C})$, $\mathcal{C}$ and $\mathcal{C}'$ are equivalent if and only if they are equivalent under a product of the following maps:

1. Independent cycling of the coordinates in each r-cycle.
2. A permutation of the f fixed points of σ.
3. A permutation of the c r-cycles of σ.
4. Multiplication by a diagonal map that is constant on each r-cycle.
5. Applying τ_u where $1 \le u < r$ to each r-cycle.

Let $\mathcal{C}$ and $\mathcal{C}'$ be equivalent with decompositions $\mathcal{C} = \mathcal{C}(\sigma) \oplus \mathcal{E}_1(\sigma) \oplus \cdots \oplus \mathcal{E}_g(\sigma)$ and $\mathcal{C}' = \mathcal{C}'(\sigma) \oplus \mathcal{E}'_1(\sigma) \oplus \cdots \oplus \mathcal{E}'_g(\sigma)$. Suppose $M = M_1 \cdots M_5$ is a product of such maps, where M_i satisfies form i above and $\mathcal{C}M = \mathcal{C}'$. Then M induces an action sending $\mathcal{C}(\sigma)^*$ to $\mathcal{C}'(\sigma)^*$ and each $\mathcal{E}_i(\sigma)^*$ to some $\mathcal{E}'_j(\sigma)^*$. In particular, $\mathcal{C}(\sigma)^*$ and $\mathcal{C}'(\sigma)^*$ are equivalent. If we then replace $\mathcal{C}'$ by an equivalent code so that $\mathcal{C}(\sigma)^*$ and $\mathcal{C}'(\sigma)^*$ are actually equal, then M must induce an automorphism of $\mathcal{C}(\sigma)^*$. Since M_1 and M_5 act trivially on $\mathcal{C}(\sigma)^*$, $M_2M_3M_4$ must induce an automorphism of $\mathcal{C}(\sigma)^*$ that sends r-cycles to r-cycles and fixed points to fixed points, putting further restrictions on the forms that need be considered for equivalence. We remark that the condition that $\langle\sigma\rangle$ be a Sylow r-subgroup of $\mathrm{Aut}(\mathcal{C})$ is satisfied for the codes of length 24 considered next.

3 Type II codes of Length 24

Using the theory of the previous two sections, we consider certain self-dual $\mathbb{Z}_4$-codes of length 24. Let $v \in \mathbb{Z}_4^n$ where $n_{\pm 1}$ components of v equal ± 1 and n_2 components of v equal 2. The *Lee weight* of v is the integer $n_{\pm 1} + 2n_2$, and the *Euclidean weight* $\mathrm{wt}_E(v)$ of v is the integer $n_{\pm 1} + 4n_2$. We remark that the Lee weight distribution is the Hamming distribution of the binary Gray map image of a $\mathbb{Z}_4$-code. A self-dual $\mathbb{Z}_4$-code $\mathcal{C}$ is of *Type II* provided it contains a vector with all entries ± 1 and $8 \mid \mathrm{wt}_E(c)$ for all $c \in \mathcal{C}$. Type II codes exist only for lengths a multiple of 8. There are 4 inequivalent Type II codes of length 8 all with minimum nonzero Euclidean weight 8 [4]. There are 133 inequivalent Type II codes of length 16 all with minimum nonzero Euclidean weight 8 [5]. A complete enumeration of such codes of length 24 seems infeasible.

In this section we discuss the self-dual $\mathbb{Z}_4$-codes of length 24 with minimum nonzero Euclidean weight 16, the highest possible minimum Euclidean weight. All such codes which have an automorphism σ of odd prime order $r \geq 5$ have been found. By the discussion in Section 2, we may assume σ has c r-cycles and f fixed points and form (2). The following table gives the numerical results, where 'number' is the number of inequivalent codes with that automorphism.

r	c	f	number
23	1	1	1
11	2	2	6
7	3	3	9
5	4	4	28

These codes were obtained by finding the possible inequivalent forms for $\mathcal{C}(\sigma)^*$ (one form occurred for each of $r = 23$, 11, and 7, and two forms for $r = 5$). Then the possible forms for $\mathcal{E}_i(\sigma)^*$ were obtained with possibilities reduced to inequivalent forms using the maps of Section 2. The minimum Euclidean weight was then found for each possibility. Details of the proof, generator matrices, and Euclidean and Lee distributions can be found in [10].

There are a total of 44 codes listed. The inequivalence of codes with the same value of r is of course settled by the theory of Section 2. It is possible that two codes arising from different values of r are equivalent. In particular, the unique code obtained when $r = 23$, which is the extended quadratic residue $\mathbb{Z}_4$-code of [1], is equivalent to one of the codes obtained when $r = 11$. In addition, this code also has minimum nonzero Lee weight 12, which is the highest minimum Lee weight of all the codes found. Another of the codes obtained with $r = 11$ is equivalent to a code with $r = 5$. There are no other equivalences among the 44 codes. In all but one pair of codes, this inequivalence between codes arising from different values of r is verified by computing the Lee and Euclidean distributions; in the exceptional case, a group theoretic argument verifies this. It is worth noting that, often, inequivalent codes arising from the same r have the same distributions, showing the value of the theory in the previous section. Indeed the 44 codes have 23 different Lee (and Euclidean) weight distributions. This gives the following result.

Theorem 3.1 *There are* 42 *inequivalent Type II codes of minimum nonzero Euclidean weight* 16 *and automorphism of odd prime order* $r \geq 5$.

We exhibit two examples.

Example 3.2 We present the two equivalent forms of the quadratic residue $\mathbb{Z}_4$-code $\mathcal{Q}_{24}$ mentioned previously. When $r = 11$ and $r = 23$, let σ_r have form (2) and let $\mathcal{C}_r$ be the associated form of $\mathcal{Q}_{24}$. When $r = 11$, $\mathfrak{R}_{11} = \mathcal{I}_0 \oplus \mathcal{I}_1$ where $\mathcal{I}_1 \simeq \mathrm{GR}(4^{10})$ has primitive element $\xi_{11} = 00021311323$ and identity $\xi_{11}^0 = 21111111111$. $\mathcal{C}_{11} = \mathcal{C}(\sigma_{11}) \oplus \mathcal{E}_1(\sigma_{11})$ is determined by

$$\mathrm{gen}(\mathcal{C}(\sigma_{11})^*) = \left[\begin{array}{cc|cc} 1 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{array}\right] \quad \text{and} \quad \mathrm{gen}(\mathcal{E}_1(\sigma_{11})^*) = \left[\begin{array}{cc} \xi_{11}^0 & \xi_{11}^{31} + \xi_{11}^{104} \end{array}\right]$$

where $\xi_{11}^{31} + \xi_{11}^{104} = 10133022121$. When $r = 23$, $\mathfrak{R}_{23} = \mathcal{I}_0 \oplus \mathcal{I}_1 \oplus \mathcal{I}_2$ where $\mathcal{I}_1 \simeq \mathcal{I}_2 \simeq \mathrm{GR}(4^1 0)$ with $\mathcal{I}_1$ having identity $\xi_{23}^0 = 12222323223322332323333$. $\mathcal{C}_{23} = \mathcal{C}(\sigma_{23}) \oplus \mathcal{E}_1(\sigma_{23}) \oplus \mathcal{E}_2(\sigma_{23})$ is determined by

$$\mathrm{gen}(\mathcal{C}(\sigma_{23})^*) = \left[\begin{array}{c|c} 1 & 1 \end{array}\right], \quad \mathrm{gen}(\mathcal{E}_1(\sigma_{23})^*) = \left[\begin{array}{c} \xi_{23}^0 \end{array}\right], \quad \text{and} \quad \mathcal{E}_2(\sigma_{23})^* = \{0\}.$$

The map

$$(1, 24, 23, 19, 9, 5, 11, 21, 15, 22, 13, 7, 2)(3, 6, 16, 12, 14)(4, 8, 17, 10)(18, 20)\mathrm{diag}(1_{11}, 3_{13})$$

sends $\mathcal{C}_{23}$ to $\mathcal{C}_{11}$. For completeness the Lee and Euclidean distributions are given in Table 1.

Example 3.3 As described earlier, there is a code $\mathcal{K}$ which has automorphisms of order 11 and of order 5; let $\mathcal{K}_{11}$ and $\mathcal{K}_5$ be the associated forms. With the notation of Example 3.2, $\mathcal{K}_{11} = \mathcal{C}(\sigma_{11}) \oplus \mathcal{E}_1(\sigma_{11})$ where

$$\mathrm{gen}(\mathcal{C}(\sigma_{11})^*) = \left[\begin{array}{cc|cc} 1 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{array}\right] \quad \text{and} \quad \mathrm{gen}(\mathcal{E}_1(\sigma_{11})^*) = \left[\begin{array}{cc} \xi_{11}^0 & \xi_{11}^0 + \xi_{11}^{341} \end{array}\right]$$

with $\xi_{11}^0 + \xi_{11}^{341} = 01311133313$. $\mathcal{K}_5 = \mathcal{C}(\sigma_5) \oplus \mathcal{E}_1(\sigma_5)$ where

$$\mathrm{gen}(\mathcal{C}(\sigma_5)^*) = \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 1 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 0 & 3 & 2 & 1 \\ 0 & 0 & 1 & 1 & 0 & 2 & 3 & 3 \\ 0 & 0 & 0 & 2 & 1 & 1 & 3 & 1 \end{array}\right] \quad \text{and}$$

$$\mathrm{gen}(\mathcal{E}_1(\sigma_5)^*) = \left[\begin{array}{cccc} \xi_5^0 & 0 & 2\xi_5 & \xi_5^0 + 2\xi_5^{11} \\ 0 & \xi_5^0 & \xi_5^0 + 2\xi_5^{14} & 2\xi_5^4 \end{array}\right]$$

with $\xi_5^0 = 03333$, $\xi_5 = 10201$, $2\xi_5 = 20002$, $\xi_5^0 + 2\xi_5^{11} = 01313$, $\xi_5^0 + 2\xi_5^{14} = 03131$, and $2\xi_5^4 = 22000$. Now ξ_{11} is a primitive element of order 1023 in $\mathcal{I}_1 \subset \mathfrak{R}_{11}$ and ξ_{11}^{341} is a cube root of unity. As $\tau_4(\xi_{11}) = \xi_{11}^4$, $\tau_4(\xi_{11}^0 + 2\xi_{11}^{341}) = \xi_{11}^0 + 2\xi_{11}^{341}$. Therefore τ_4 applied to each 11-cycle induces an automorphism of $\mathcal{K}_{11}$ of order 5. In the list of codes associated to $r = 5$, only the code $\mathcal{K}_5$ has the same Lee and Euclidean distributions (given in Table 1) as $\mathcal{K}_{11}$ does. Hence $\mathcal{K}_{11}$ and $\mathcal{K}_5$ are equivalent.

Lee weight	$\mathcal{Q}_{24}$	$\mathcal{K}$
0, 48	1	1
8, 40	0	66
10, 38	0	0
12, 36	12144	10560
14, 34	61824	74496
16, 32	195063	127215
18, 30	1133440	1365760
20, 28	1445136	918720
22, 26	4080384	4916736
24	2921232	1950108

Euclidean weight	$\mathcal{Q}_{24}$	$\mathcal{K}$
0, 96	1	1
16	98256	97794
24	1987616	1999232
32	6546375	6511791
40	5974848	6006528
48	1925376	1928412
56	231840	213888
64	12903	19503
72	0	0
80	0	66

Table 1: Lee and Euclidean distributions of $\mathcal{Q}_{24}$ and $\mathcal{K}$.

Even unimodular lattices can be constructed from linear $\mathbb{Z}_4$-codes $\mathcal{C}$ of length n as follows. Let

$$\Lambda(\mathcal{C}) = \{x \in \mathbb{Z}^n \mid x \equiv c \pmod 4 \text{ for some } c \in \mathcal{C}\}.$$

By [1, Theorem 4.1], $\Lambda(\mathcal{C})/2$ is an even unimodular lattice whenever $\mathcal{C}$ is a self-dual $\mathbb{Z}_4$-code with all Euclidean weights divisible by 8. The 42 inequivalent codes have minimum nonzero Euclidean weight 16; hence each corresponding lattice is an even unimodular lattice of minimum norm 4 in $\mathbb{R}^{24}$. By [3, Chapter 12], these lattices are all the Leech lattice Λ_{24}.

References

[1] A. Bonnecaze, P. Solé, and A. R. Calderbank, "Quaternary quadratic residue codes and unimodular lattices," *IEEE Trans. Inform. Theory* **IT-41** (1995), 366–377.

[2] A. R. Calderbank and N. J. A. Sloane, "Modular and p-adic cyclic codes," *Designs, Codes and Cryptography* **6** (1995), 21–35.

[3] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups*, New York: Springer-Verlag, 1988.

[4] J. H. Conway and N. J. A. Sloane, "Self-dual codes over the integers modulo 4," *J. Comb. Theory A* **62** (1993), 30–45.

[5] J. Fields, J. S. Leon, and V. Pless, "All $\mathbb{Z}_4$ codes of type II and length 16 are known," *J. Comb. Theory A* **78** (1997), 32–50.

[6] A. R. Hammons Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, "The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes," *IEEE Trans. Inform. Theory* **IT-40** (1994), 301–319.

[7] W. C. Huffman, "Automorphisms of codes with applications to extremal doubly even codes of length 48," *IEEE Trans. Inform. Theory* **IT-28** (1982), 511–521.

[8] W. C. Huffman, "On the [24, 12, 10] quaternary code and binary codes with an automorphism having two cycles," *IEEE Trans. Inform. Theory* **IT-34** (1988), 486–493.

[9] W. C. Huffman, "On extremal self-dual quaternary codes of lengths 18 to 28, I," *IEEE Trans. Inform. Theory* **IT-36** (1990), 651–660.

[10] W. C. Huffman, "Decompositions and extremal type II codes over $\mathbb{Z}_4$," *IEEE Trans. Inform. Theory*, to appear.

[11] B. R. McDonald, *Finite Rings with Identity*, New York: Marcel Dekker, Inc., 1974.

[12] Z. Qian, *Cyclic Codes over* $\mathbb{Z}_4$, Doctoral Thesis, University of Illinois at Chicago, 1996.

[13] V. Y. Yorgov, "Binary self-dual codes with automorphisms of odd order," *Prob. Inform. Trans.* **XIX** (1983), 11–24.

Distance regular spherical codes

Thomas Ericson
Linköping University
Dept. of Electrical Engineering
S-581 83 Linköping, Sweden
thomas@isy.liu.se

Abstract

Distance regular spherical codes are closely related to association schemes. We develop the basic theory and illustrate with the aid of several examples.

1. Introduction

A **spherical code** is a finite subset $C = \{x_1, x_2, \ldots, x_v\}$ of unit vectors in Euclidean space $\mathbf{R}^n$. The smallest n such that $\mathbf{R}^n$ contains C is called the **dimension** of the code. The vectors x_i are called **codewords**. The number v of codewords is called the **size** of the code. A matrix X having the codewords x_i of C as its rows is called a **code matrix**; the matrix $G = X X^T$ is a **Gram matrix**. The rows x_i are allowed to occur in any order in the code matrix X, so X and $Y = PX$, where P is a permutation matrix, are different code matrices corresponding to the same code. The Gram matrix corresponding to Y is PGP^T.

The Gram matrix is square of order v, where v is the size of the code. The rank of the Gram matrix G equals the dimension n of the code.

Two codes which can be transformed to each other by an orthogonal transformation are said to be **equivalent**. The corresponding code matrices X and Y are related by an equation of the form $Y = XT$, where T is an orthogonal matrix. Two equivalent code matrices generate the same Gram matrix, while a permutation of rows changes a Gram matrix G to an equivalent matrix PGP^T, where P is a permutation matrix.

An orthogonal transformation T which transforms a code C into itself is a **symmetry** of the code. A symmetry T transforms the code matrix X to a matrix XT, such that $XT = PX$ where again P is a permutation matrix. The set of all symmetries T to a given spherical code C is denoted Symm(C). It is readily seen to be a group. It will be referred to as the **symmetry group** of the code. By mapping each matrix T in Symm(C) to the permutation matrix P defined by XT = PX we obtain a homomorphic image of Symm(C) in the form of a group of permutation matrices. This group will be referred to as the **induced permutation group**.

Each element in the Gram matrix G is an inner product (x,y) between codewords in C. Let the possible inner products be $1 = s_0 > s_1 > \ldots > s_d$. The number d of different values of the non-zero inner products is called the **degree** of the code. The corresponding squared Euclidean distances are $\rho_i = 2 - 2s_i$, i = 0, 1, . . . , d. Define

$$A_i(x, y) = \begin{cases} 1 & (x,y) = s_i \\ 0 & (x,y) \neq s_i \end{cases}$$

$$i = 0, 1, \ldots, d; \qquad x, y \in C .$$

The matrices A_i are square of order v, where v is the size of the code. The following properties are satisfied:

$$\text{i) } A_0 = I, \qquad \text{ii) } A_i^T = A_i , \qquad \text{iii) } \sum_{i=0}^{d} A_i = J ,$$

where I denotes the v x v identity martix and J denotes the v x v matrix with ones in all positions. It follows that the matrices A_i can be regarded as adjacency matrices to graphs Γ_i, with the code C as a common vertex set. Two vertices x, y in C are joined by an edge in Γ_i if and only if (x, y) = s_i . The graph Γ_0 corresponding to the matrix $A_0 = I$ is the empty graph. The remaining graphs Γ_i , i = 1, 2, . . . , d form a partition of the complete graph K_v over the vertex set C . The graphs Γ_i will be referred to as the **distance graphs** of the code C .

Recall that a graph is said to be regular if each vertex is joined to the same number of edges. The number of edges connected to each vertex in a regular graph is called the **valency** of the graph. We say that a spherical code is **regular** if all its distance graphs are regular in the sense of graphs. Let C be a regular spherical code and let the valencies of its distance graphs be k_i, i = 1, 2, . . . , d. We define separately $k_0 = 1$. It follows that each codeword x in C has k_i codewords on squared distance ρ_i . The sequence $\{(\rho_i,k_i): 0 \le i \le d\}$ is called the **distance distribution**. Notice that this concept is defined only if the code is regular.

2. Group codes

The symmerty group Symm(C) of a spherical code C is a group of orthogonal matrices operating on the Euclidean space $\mathbf{R}^n$. The code is a union of orbits. A spherical code consisting of one single orbit is called a **group code**. The concept was introduced by Slepian (1968), [1], who noticed that many practically important codes are group codes. It has later been studied by Forney (1991), [2] under the name of **geometrically uniform** codes.

Any group code is regular. The converse is not true: a regular spherical code need not be a group code. Slepian showed this by providing a counter-example. We will demonstrate the same fact using a slightly different argument developed by Spets (1997), [3].

A sequence $x_1, x_2, \ldots, x_m$ of codewords from a given spherical code C is said to be a **distance cycle** if the following holds:

$$(x_k, x_{k+1}) = (x_1, x_m), \quad k = 1, 2, \ldots, m-1 .$$

The number m is called the **length** of the cycle. When regarded as vertices in the graph Γ_i generated by the inner product $s_i = (x_1, x_m)$ the codewords $x_1, x_2, \ldots, x_m$ form a cycle in the sense of graphs. The following theorem is true.

Theorem: Let C be a group code of size v and let p be a prime dividing v. Then the code C contains a distance cycle of length p.

Proof: As C is an orbit under Symm(C) and as p divides the size v of C it follows that p divides also the order of Symm(C) . By Cayley's theorem there is an element T in Symm(C) having order p. Let x be an arbitrary codeword in C . Define $x_i = x\,T^i$, i = 0, 1, . . . , p-1. It is clear that $\{x_i: 0 \le i \le p-1\}$ is a distance cycle of length p.

With the aid of this theorem it is easy to construct examples of regular codes which are not group codes. We first observe that the induced permutation group is a subgroup in the automorphism group of each one of the graphs Γ_i . Thus, if for some prime p we can find a partition $\{\Gamma_i : 1 \le i \le d\}$ of the complete graph K_v such that all of the graphs Γ_i are regular and none of them contains a cycle of order p , then it follows that any spherical code having $\{\Gamma_i : 1 \le i \le d\}$ as its distance graphs is regular without being a group code. In Fig.1 a simple example of this kind is given. It is easy to construct a code having these graphs as their distance graphs. In fact, it is possible to construct a sequence of such codes converging to the simplex code in dimension n = 6.

3. Distance regular spherical codes

Strengthening the regularity condition a little bit further we arrive at the concept of distance regular spherical code. The (x,y) element in the matrix product A_iA_j indicates the number of codewords that are simultaneously on squared distance ρ_i from x and on squared distance ρ_j from y. In general this number depends on the pair {x,y}. A code with the property that it depends only on the **distance** between x and y is said to be **distance regular**.

It is clear that any such code is regular. The converse is not true. A simple counter-example is indicated in Fig.2.

Let $A_iA_j(x,y) = p_{ij}^k$ for $(x,y) = s_k$, $k = 0, 1, \ldots, d$. It follows that the product A_iA_j has an expansion in the form

$$\text{iv)} \qquad A_iA_j = \sum_{k=0}^{d} p_{ij}^k A_k .$$

As the matrices A_i are symmetric the following property also follows:

$$\text{v)} \qquad A_iA_j = A_jA_i .$$

Let $\mathcal{A} = \{A_0, A_1, \ldots, A_d\}$ be a set of indicator matrices satisfying the properties i) - v). Consider the linear space $L(\mathcal{A})$ spanned by the matrices A_i in $\mathcal{A}$. It follows that this space is a commutative algebra. In fact, the properties above are exactly those of an **association scheme**. Thus, any distance regular spherical code induces an association scheme. Even more: also the converse is true.

The algebra $L(\mathcal{A})$ is usually called the **Bose-Mesner algebra**. As each matrix A in $L(\mathcal{A})$ is symmetric it follows that the eigenvalues are real. As the algebra is commutative it follows that all elements in $L(\mathcal{A})$ are jointly diagonalized by a fixed orthogonal matrix. The different eigenvalues correspond to different eigenspaces and these eigenspaces are common for all elements in $L(\mathcal{A})$.

Let A be an arbitrary matrix in the Bose-Mesner algebra and let U_θ be a matrix formed by a set of unit length mutually orthogonal column vectors spanning the eigenspace corresponding to the eigenvalue θ . We have

$$AU_\theta = \theta\, U_\theta \,.$$

The matrix U_θ has v rows and m columns, where m is the multiplicity of the eigenvalue θ. The matrix $E_\theta = U_\theta U_\theta^T$ is square of order v. It enjoys a number of useful properties (Godsil [4], p 28) :

i) $E_\theta^2 = E_\theta$

ii) $AE_\theta = \theta E_\theta$

iii) $\sum_{\theta \in ev(A)} E_\theta = I$

iv) $p(A) = \sum_{\theta \in ev(A)} p(\theta)\, E_\theta \,,$ (p(x) is a polynomial)

where ev(A) denotes the set of all eigenvalues of the matrix A.

Each one of the matrices E_θ can be expressed as a polynomial in the matrix A and so belongs to the Bose-Mesner algebra. To see this, define

$$f(x) = \prod_{\tau \in ev(A) \setminus \theta} (x - \tau)\,.$$

By iv) we have

$$f(A) = \sum_{\tau \in ev(A)} f(\tau)\, E_\tau = f(\theta)\, E_\theta \,.$$

By construction $f(\theta) \neq 0$, and so we obtain

$$E_\theta = \frac{1}{f(\theta)}\, f(A).$$

The matrix E_θ is obviously positive semidefinite. In addition it is constant along the main diagonal. To see this, notice that f(A) - being an element in the Bose-Mesner algebra - has an expansion in terms of the matrices A_i . Let the expansion be

$$f(A) = \sum_{i=0}^{d} f_i\, A_i \,.$$

The statement follows because all matrices A_i have zero diagonals, except for $A_0 = I$. As a consequence all diagonal elements in the matrix E_θ have the same value, namely $f_0/f(\theta)$. This in turn means that the rows in U_θ have constant energy. Defining

$$X = \sqrt{\frac{f(\theta)}{f_0}}\ U_\theta$$

we obtain a matrix in which all the rows have unit energy. Some rows might coincide, but the set of different rows defines a spherical code. The Gram matrix is given by

$$G = XX^T = \frac{f(\theta)}{f_0}\ U_\theta U_\theta{}^T = \sum_{i=0}^{d} \frac{f_i}{f_0} A_i \,.$$

4. Examples

A main goal in connection with spherical codes is to find codes with large minimum distance given the size and the dimension. Although distance regular codes do not always have good properties in this sense there are many examples where they do. In three dimensions all the Platonic solids are distance regular. Of these the tetrahedron, the octahedron and the icosahedron are optimal in the sense of maximizing the minimum distance. The spherical code obtained by the binary antipodal representation of the binary parity check code of lenght 5 is an optimal spherical code of size $v = 16$, with squared minimum distance $\rho = 8/5$. It is distance regular. The inner shell of the D_4 lattice is a spherical code of size $v = 24$ and squared minimum distance $\rho = 1$. It is a best known solution for contact number. It is distance regular.

In dimension $n = 3$ the snub cube with size $v = 24$ and the square antiprism with size $v = 8$ offer examples of optimal spherical codes that are regular without being distance regular.

A distance regular spherical code with degree $d = 2$ has two distance graphs. Both are examples of so called **strongly regular graphs**. Such graphs have been extensively studied in the literature. An excellent survey can be found in Cameron-van Lint, [5]. Any strongly regular graph generates a pair of distance regular spherical codes. Many of the best spherical codes are generated by such graphs. In Table 1 we have listed a few examples. References are given to the standard text [5].

References

[1] D. Slepian, Group codes for the Gaussian channel, Bell Syst.Techn. J., vol. 47, pp 575-602, April 1968.

[2] G.D. Forney Jr, Geometrically uniform codes, IEEE Trans. Inform. Theory, vol .37, pp1241-1260, Sept. 1991.

[3] E. Spets, Geometrically uniform sets, NTNU Trondheim, Norges teknisk-naturvetenskapliga institut, Thesis 1997:31

[4] C.D. Godsil, Algebraic combinatorics, Chapman and Hall, 1993.

[5] P.J. Cameron and J.H. van Lint, Designs, Graphs, Codes and their links, London Mathematical Society, Student Texts 22, 1991.

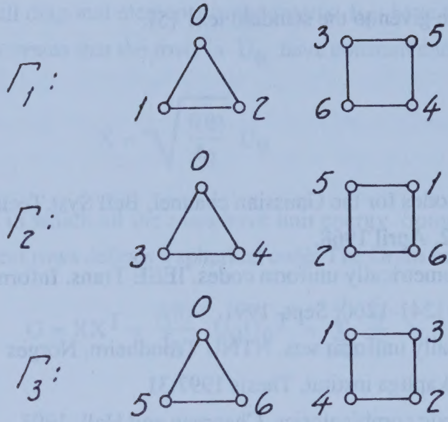

Γ1:
0
1
2
3
5
6
4
Γ2:
0
3
4
5
1
2
6
Γ3:
0
5
6
1
3
4
2

Fig.1. Distance graphs for a regular spherical code which is **not** a group code (from E.Spets,[3], p 62).

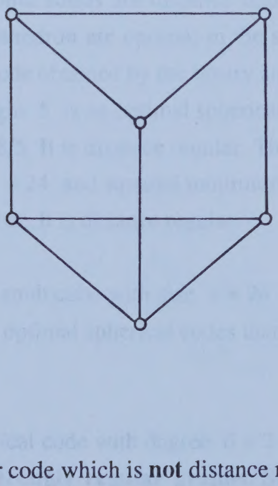


Fig.2. A regular code which is **not** distance regular.

v	k	a	c	n	ρ	M	Reference
10	3	0	1	4	1.67	10	Petersen, p 33.
16	5	0	2	5	1.60	16	Clebsch, p 35.
27	10	1	5	6	1.50	27	Schläfi, p32.
56	10	0	2	20	1.87	56	Gewirtz, p 32.
120	42	8	18	20	1.71	120	D^{pq}, p 72.
50	7	0	1	21	1.90	50	Hoffman-Singleton
77	16	0	4	21	1.83	77	D_p, p 72.
112	30	2	10	21	1.78	112	1st subconst. McL
162	56	10	24	21	1.71	162	2nd subconst. McL
176	70	18	34	21	1.67	176	D^p, p 72.
100	22	0	6	22	1.82	100	Higman-Sims, p 109.
253	112	36	60	22	1.64	253	D, p 72.
275	112	30	56	22	1.67	275	McLaughlin, p 112.

Table 1. Spherical codes generated from regular graphs. The first four columns give the parameters of the graph ((n, k, λ, μ) in the notation of Cameron-van Lint, [5]). The next three columns give the parameters of the code: dimension n, squared minimum distance ρ, and size M. References are given to Cameron-van Lint, [5].

SPLIT WEIGHT ENUMERATORS OF EXTREMAL SELF-DUAL CODES

JOE FIELDS, VERA PLESS

ABSTRACT. The split weight enumerator has been used to aid in the determination of the existence of a code with a particular weight enumerator. We have found a new faster method of computing split weight enumerators of self-dual codes using specific properties of these codes. These techniques are applied to the weight enumerators of putative extremal [72,36,16] and [96,48,20] codes and other interesting codes.

1. INTRODUCTION

A doubly-even, self-dual binary block code must have length n a multiple of 8. Such a code can have minimum distance at most $4+4\lfloor n/24 \rfloor$ and codes having this minimal distance are called extremal. Thus, at lengths 8 and 16, we have extremal codes having minimum distance 4 (the length 8 extended Hamming code e_8 and two codes of length 16: d_{16} and $e_8 \oplus e_8$). At length 24 the remarkable extended Golay code is the unique extremal doubly-even code. At length 32 there are 5 inequivalent extremal doubly-even codes [3], and this is the largest length where the complete story is known. Conway and Pless determined these 5 inequivalent extremal codes by completely classifying all of the doubly-even self-dual codes of length 32. This approach appears to be impractical for lengths greater than 32. There is one known extremal code of length 48 (an extended quadratic residue code) and it is an open question whether there are others. The smallest length for which the existence of an extremal code is not known is 72.

The MacWilliams relations, which are a set of equations relating the weight enumerator of a code to that of its dual, often serve to completely determine the weight enumerator of a self-dual code. Indeed if the code is doubly-even and extremal, the MacWilliams relations uniquely determine the weight enumerator.

Split weight enumerators are useful in investigating whether there can be a code with a certain weight enumerator. The split weight enumerator has to exist with non-negative integer coefficients. As we shall see, it often implies the existence of other smaller codes with certain weight enumerators. Split weight enumerators can also be used to determine the weight enumerator of certain cosets of a code.

When the usual system of equations (1) are used, the calculation of the split weight enumerator takes a great deal of time. With other methods we have been able to determine the split weight enumerators of certain interesting doubly-even self-dual codes of length > 100 which were not possible to find using (1). We describe these methods and give specific results for certain doubly-even self-dual codes.

2. PRELIMINARIES

The weight enumerator of a code is a polynomial in one variable (say x) such that the coefficient of x^i is the number of codewords having Hamming weight i. The split weight enumerator of a code with respect to a partition is a polynomial in several variables, one for each piece of the partition. For instance, if we consider the famous extended Golay code and a partition of the coordinates into two pieces: the support of a minimum weight codeword in the first part, and its complement, in the second part, we get a polynomial in two variables. The coefficient of x^iy^j is the number of words in the code having i ones in the first part of the partition and j ones in the second.

The split weight enumerator of the Golay code (partitioned as above) is:

$$1 + x^8 + 280x^4y^4 + 448x^2y^6 + 30y^8 + 448x^6y^6 + 1680x^4y^8$$
$$+ 448x^2y^{10} + 30x^8y^8 + 448x^6y^{10} + 280x^4y^{12} + y^{16} + x^8y^{16}$$

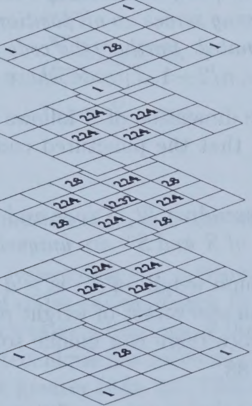
The symmetries of the situation may be more obvious if we instead display this information as a table (the number in the (i, j) position is the coefficient of x^jy^i in the split weight enumerator.

For compactness, we display only the even-numbered columns and rows:

	$j = 0$	2	4	6	8
$i = 0$	1	0	0	0	1
2	0	0	0	0	0
4	0	0	280	0	0
6	0	448	0	448	0
8	30	0	1680	0	30
10	0	448	0	448	0
12	0	0	280	0	0
14	0	0	0	0	0
16	0	1	0	0	1

We will usually denote the number in the (i, j) position of such a table by $X_{i,j}$.

Next, let us consider a partition of the coordinates into three parts, each the support of an octad. The resulting numbers could be expressed as the coefficients of a polynomial in three variables or (as we shall do) as a three-dimensional array of numbers:



In [11], Simonis introduced the generalized MacWilliams identities which can be used to find these split weight distributions. Suppose that T is a partition of n into p parts. $W(T)$ is the set of all possible weight vectors with respect to the partition, i.e., $W(T) = \{\mathbf{i} \in \mathbb{N}^p \mid i_u \leq n_u \text{ for all } u \text{ with } 1 \leq u \leq p\}$, where n_u is the size of the u-th piece of the partition T, and i_u is the u-th coordinate of the vector $\mathbf{i}$.

For some codeword $\mathbf{x} \in C$, the T-weight, or weight with respect to the partition, is a vector of non-negative integers giving the weight of $\mathbf{x}$ in each piece of the partition. Denote the T-weight of $\mathbf{x}$ by $w_T(\mathbf{x})$. Clearly $w_T(\mathbf{x})_u \leq n_u$, for all $1 \leq u \leq p$. Define $A_{\mathbf{i}}(T) = \{\mathbf{x} \in C \mid w_T(\mathbf{x}) = \mathbf{i}\}$ and similarly, define $B_{\mathbf{i}}(T) = \{\mathbf{y} \in C^\perp \mid w_T(\mathbf{y}) = \mathbf{i}\}$. Note that $|A_{\mathbf{i}}(T)|$ (resp. $|B_{\mathbf{i}}(T)|$) is one component of the split weight distribution of C (resp. $C^{\perp}$).

For binary codes, the generalized MacWilliams identities are as follows.

$$(1) \qquad \sum_{\mathbf{j}\in W(T)} \left[\prod_{u=1}^{p} \binom{n_u - j_u}{i_u}\right] A_{\mathbf{j}}(T) = 2^{k-\|\mathbf{i}\|} \sum_{\mathbf{j}\in W(T)} \left[\prod_{u=1}^{p} \binom{n_u - j_u}{n_u - i_u}\right] B_{\mathbf{j}}(T)$$

Note that there is an equation for each (vector) $\mathbf{i}$ in $W(T)$.

For even the most trivial examples, the generalized MacWilliams identities give us extremely large systems of equations involving an equally large number of unknowns. For instance, consider the $[8,4,4]$ extended Hamming code and a partition of its coordinates into two pieces: the support of disjoint codewords. The split weight enumerator (again only the even columns and rows are shown) is:

	$j = 0$	2	4
$i = 0$	1	0	1
2	0	12	0
4	1	0	1

But the generalized MacWilliams identities give a system of 25 equations in 25 unknowns. By naively sending these equations to a linear system solver we find the unique solution, but by first taking into account some of the facts which are known about this code (that it is doubly-even and has minimum weight 4) we get a much smaller system. Indeed, if we also allow ourselves to note that the 4 corner entries must be 1's, we get a system that can be solved merely by considering the dimension of the code!

If we are interested in the split weight enumerator of a code C relative to a partition of its coordinates into two parts, where one portion holds a vector of weight w in C, then knowledge of the shortened code in the other portion is useful.

Theorem 1. *If C is an $[n, n/2, d]$ self-dual binary code containing a vector v of weight $w < 2d$, the subcode of C, S, having zeroes on all positions in v is an $[n-w, n/2-w+1, d]$ code if those positions are eliminated. Further, the code derived from C by puncturing on the coordinates of v is an $[n-w, n/2-1, d']$ code where $d' \geq d - w/2$ and, in fact, is $S^{\perp}$.*

Proof. The statement about the dimension of S follows from Theorem 103 in [10] or the balance principle [6]. The fact that the punctured code is $S^{\perp}$ is also in Theorem 103 [10]. □

Theorem 2. *If C is a doubly-even $[n, n/2, d]$ code and v is a vector in C of weight $w < 2d$, then the weight distribution of S and $S^{\perp}$ are uniquely determined iff $w \leq 6d - n - 4$.*

Proof. We know that the minimum weight of $S^{\perp}$ is $\geq d - w/2$ and the minimum weight of S is d. Now S contains the all one vector of weight $n - w$ so the number of unknown weights in S is $\frac{n-w-2d}{4} + 1$. Hence there is a unique solution to the power moments iff $\frac{d-w}{2} \geq \frac{n-w-2d}{4} + 1$ [10]Theorem 83. □

Note that this theorem implies that a unique solution for the weight enumerators of S and $S^{\perp}$ is only possible for extremal doubly-even codes and then only for $w \leq 20$.

If we consider the split weight enumerator of a doubly-even $[n, n/2, d]$ code with respect to a vector v of weight w, then the weight distribution of S gives all the terms in this split weight enumerator of the form $X_{i,0}$. The terms in the weight enumerator of $S^{\perp}$ are related to a sum of terms in the split weight enumerator, i.e. if b_i is the number of vectors of weight i in $S^{\perp}$, $2b_i = \sum_j X_{i,j}$. The factor of 2 comes from the fact that for each codeword $x \in C$ that is mapped to a codeword $y \in S^{\perp}$, the vector $x + v$ is also. That is, there is a 2 to 1 correspondence between C and $S^{\perp}$.

Essentially, this last paragraph says that the weight distribution of S is in the far left (and far right) column of the split weight distribution table and that by summing rows of this table we get twice the weight distribution of $S^\perp$.

We also get known results if we sum the split weight distribution table along columns and diagonals.

The diagonal sums are perhaps easiest to see. If we sum along the diagonals such that $i + j$ is constant, we will find all the vectors of a fixed weight in the original code C. Thus $A_k = \sum_{i+j=k} X_{i,j}$.

Since S may be viewed as a subcode of C, by summing the columns we find all the vectors in cosets of S with a fixed weight coset representative. Thus $\sum_i X_{i,j} = \binom{w}{j} 2^{n-w-k+1}$.

We use the equations given in the last several paragraphs to define a linear system of equations whose unknowns are the variables $X_{i,j}$ in the split weight distribution. This system is, in general, underdetermined. If we add some additional assumptions about the code (that it is doubly-even and self-orthogonal), and the fact that the split weight distribution itself has certain symmetries, we sometimes get a determinate system. This technique was sufficient to find the split weight distributions of all extremal doubly-even self-dual codes for lengths less than 72.

The complete system of equations is:

$$X_{i,j} = 0 \text{ if } i + j \not\equiv 0 \pmod 4 \text{ or } i \not\equiv 0 \pmod 2$$
$$X_{i,j} = X_{i,w-j} = X_{n-w-i,j} = X_{n-w-i,w-j}$$
$$\sum_i X_{i,j} = \binom{w}{j} 2^{n-w-k+1}$$
$$\sum_{i+j=k} X_{i,j} = A_k$$
$$\sum_j X_{i,j} = 2b_i$$
$$X_{0,i} = X_{w,i} = a_i$$

Where A_k is the the weight distribution of C, a_i is the weight distribution of S and b_i is the weight distribution of $S^\perp$.

As was mentioned above, this system was not sufficient to determine the split weight distributions of putative extremal codes having length 72 or greater. For these codes, we developed another technique which takes advantage of the fact that vectors of all weights in such codes hold designs. This technique has its own limitations – it cannot be applied to codes having minimum weight greater than 20. This seems to be related to but not a direct consequence of Theorem 2. We are able to determine split weight distributions in terms of a single parameter for some extremal codes partitioned into the support of a weight 24 vector and its complement.

The design technique consists of generating a different linear system, the equations come from counting (in two ways) flags of the form: sets of points (of size $\leq t$) from the first part of the partition contained in a block of the t-design held by vectors of a fixed weight.

For instance, consider the putative $[72, 36, 16]$ code and a partition whose first part is the support of a minimum weight vector. The equations determined by the designs held by weight 16 vectors are:

$$\binom{16}{1}(\lambda_{1,1}-1)=\binom{2}{1}X_{14,2}+\binom{4}{1}X_{12,4}\binom{6}{1}X_{10,6}+\binom{8}{1}X_{8,8}$$

$$\binom{16}{2}(\lambda_{2,2}-1)=\binom{2}{2}X_{14,2}+\binom{4}{2}X_{12,4}\binom{6}{2}X_{10,6}+\binom{8}{2}X_{8,8}$$

$$\binom{16}{3}(\lambda_{3,3}-1)=\binom{4}{3}X_{12,4}\binom{6}{3}X_{10,6}+\binom{8}{3}X_{8,8}$$

$$\binom{16}{4}(\lambda_{4,4}-1)=\binom{4}{4}X_{12,4}\binom{6}{4}X_{10,6}+\binom{8}{4}X_{8,8}$$

$$\binom{16}{5}(\lambda_{5,5}-1)=\binom{6}{5}X_{10,6}+\binom{8}{5}X_{8,8}$$

There are five equations, because the weight 16 vectors hold a 5-design, thus the first equation reflects the fact that we have a 1-design, the second a 2-design, etc.

Since there can only be 5 equations, this system can (at best) be used to discover the values of 5 unknowns. We express the coefficients of that (slightly larger) system in matrix form.

$$\begin{bmatrix} \binom{2}{1} & \binom{4}{1} & \binom{6}{1} & \binom{8}{1} & \binom{10}{1} \\ \binom{2}{2} & \binom{4}{2} & \binom{6}{2} & \binom{8}{2} & \binom{10}{2} \\ 0 & \binom{4}{3} & \binom{6}{3} & \binom{8}{3} & \binom{10}{3} \\ 0 & \binom{4}{4} & \binom{6}{4} & \binom{8}{4} & \binom{10}{4} \\ 0 & 0 & \binom{6}{5} & \binom{8}{5} & \binom{10}{5} \end{bmatrix}$$

It can easily be checked that this system is invertible (its determinant is 32768).

Fact 1. *Let C be a doubly-even $[n, n/2, d]$ code and consider its split weight enumerator with respect to a vector v in C of weight $w \leq 2d$. Then if the weight distribution of S and $S^{\perp}$ are uniquely determined, so is the split weight enumerator.*

This is so by observation of the finite number of cases where this can occur, which are given in the next section.

For all the cases where we have seen that the weight enumerators of S and $S^{\perp}$ can be given in terms of one variable, the split weight enumerator can also be given in terms of one variable.

3. THE CASES

We have computed split weight enumerators for many extremal doubly-even codes. As we see from Theorem 2, the weight distributions of S and $S^{\perp}$ are uniquely determined whenever we choose the codeword v of weight ≤ 20. We also determined the split weight enumerators for v of weight 24 where the weight distributions of S and $S^{\perp}$ (and also, the split weight enumerator) can be given in terms of one variable. We discuss these first. The other case of interest is the extended code of the code generated by the lines of a projective plane of order 10 (notwithstanding the fact that this plane is thought not to exist!)

3.1. Extremal Doubly-even Codes.

$n = 24$) There are unique split weight enumerators with respect to words of weights 8 and 12.

$n = 32$) There are unique split weight enumerators with respect to words of weights 8 and 12. There is a one-variable solution when partitioning on the support of a weight 16 codeword. Note that there are 5 inequivalent $[32, 16, 8]$ doubly-even codes [3]. There are 8 different split weight enumerators if the partition is into two portions of size 4 and 28; clearly neither of these is the support of a codeword. These correspond to the different possible weight distributions for cosets of weight 4. Cosets of weights 1,2,3,5 and 6 have unique distributions [1].

$n = 48$) There are unique split weight enumerators with respect to words of weights 12, 16 and 20. There is a one-variable solution for weight 24. Here it is not known if there is any other code than the extended quadratic residue code [5]. It is known that the weight distributions of cosets of weights 1,2,3,4,5,7 and 8 are unique [1]. There are 7 possible weight distributions for cosets of weight 6. Six of them appear in the weight 6 cosets of the extended quadratic residue code.

$n = 72$) There are unique split weight enumerators with respect to words of weights 16 and 20. For both of these partitions, there are known codes having the parameters of S and $S^{\perp}$ [2] but these codes do not have the weight distributions dictated by the split weight enumerators. For example, shortening a $[72, 36, 16]$ code on the support of a weight 16 vector gives a code S having parameters $[56, 21, 16]$. Such a code is known [2], and it turns out to be doubly-even, but its weight distribution is

$$[1, 0, 0, 0, 0, 0, 0, 0, 5170, 0, 90640, 0, 508365, 0, 888800,$$
$$0, 508365, 0, 90640, 0, 5170, 0, 0, 0, 0, 0, 0, 0, 1]$$

whereas the weight distribution of S must be

$$[1, 0, 0, 0, 0, 0, 0, 0, 5082, 0, 91168, 0, 507045, 0, 890560,$$
$$0, 507045, 0, 91168, 0, 5082, 0, 0, 0, 0, 0, 0, 0, 1].$$

There is a one-variable solution for weight 24.

$n = 96$) There is a unique split weight enumerator with respect to words of weight 20. Here, S is a $[76, 29, 20]$ code. The best known $[76, 29]$ code has minimum weight 18 [2], on the other hand, the tightest upper bound on the minimum weight of such a code is 22 [2]. In contrast, there is a known code having parameters $[76, 47, 10]$ (those of $S^{\perp}$) [2] but this code does not have the weight distribution dictated by the split weight enumerator.

There is a one-variable solution for weight 24.

$n = 120$) There is a one-variable solution for weight 24.

3.2. **The Projective Plane of Order 10.** Let C be the $[111, 56, 11]$ code generated by vectors of a projective plane of order 10. It is known that the only vectors of weight 11 are the 111 lines of the plane. It was shown [7, 8] that C contains no vectors of weights 12,15 or 16. At this point the weight distribution of C was uniquely determined. It is not hard to see that a $[111, 56, 11]$ code with $A_{11} = 111$, $A_i = 0$ unless $i = 3, 0 \pmod 4$ and $A_{12} = A_{15} = A_{16} = 0$ must be the code of a plane of order 10. The sum of any two vectors of weight 11 can only have weight 20 so they must meet in exactly one point. Hence at most 11 vectors of weight 11 can contain the same point. There must be 11 if we count the number of 1's in the matrix M of vectors of weight 11 by columns. From this it follows that any 2 points must be on exactly one line.

Let $\overline{C}$ be the $[112, 56, 12]$ doubly-even extended code of C, then $\overline{A}_{12} = 111$ and $\overline{A}_{16} = 0$. Note that any two vectors of weight 12 meet (indeed they meet in two points – the same point the corresponding weight 11 vectors met and the extended point). Lam [9] showed that there is no projective plane of order 10 by demonstrating, with computers running for a year, that C could not contain any vectors of weight 19. This was a contradiction as the unique solution for the weight distribution of C had $A_{19} \neq 0$.

We examined $\overline{C}$ with the stated assumptions. We consider the split weight distribution with respect to a vector of weight 12. We let S be the $[100, 45, 20]$ shortened code and $S^{\perp}$ be the $[100, 55, 10]$ punctured code. Let b_i denote the number of vectors of weight i in $S^{\perp}$. All vectors in $S^{\perp}$ have even weight. As $b_{10} = 110$ and $b_{12} = 0$ and S contains vectors of 16 unknown weights, it is possible to find the weight distributions of S and $S^{\perp}$ in terms of one variable. The split weight enumerator can be computed and expressed in terms of one variable as well.

4. Appendix: Split weight distributions of selected codes

Since these split weight distribution tables are symmetrical around both their horizontal and vertical midpoints, we only display one quarter of the tables.

The putative $[72, 36, 16]$ doubly-even self-dual extremal code partitioned into two parts, one being the support of a minimum weight vector.

1	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	2310
0	0	0	34496	0
0	0	123480	0	831600
0	84480	0	5396160	0
5082	0	8796480	0	62455470
0	2956800	0	198007040	0
91168	0	166529440	0	1176806400
0	29944320	0	1997318400	0
507045	0	922595520	0	6525588300
0	92843520	0	6196240512	0
890560	0	1620726800	0	11458978080

The putative $[72, 36, 16]$ doubly-even self-dual extremal code partitioned into two parts, one being the support of a weight 20 vector.

1	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	476
0	0	0	0	11475	0
0	0	0	74460	0	338096
0	0	121380	0	3018435	0
0	40800	0	7956000	0	38079660
1258	0	5827260	0	152023860	0
0	940800	0	192637200	0	917613760
14637	0	71092980	0	1847582100	0
0	5986720	0	1220548320	0	5819294072
49640	0	240480300	0	6252934050	0
0	10967040	0	2237918760	0	10665686304

The putative [72, 36, 16] doubly-even self-dual extremal code partitioned into two parts, one being the support of a weight 24 vector. This distribution is not unique, but can be expressed in terms of a single variable.

1	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	2z
0	0	0	0	0	3312-12z	0
0	0	0	0	32706+30z	0	115920+64z
0	0	0	108192-40z	0	1493712-132z	0
0	0	88872+30z	0	5941728+120z	0	174z+21944944
0	16560-12z	0	7534800-20z	0	110157120+48z	0
207+2z	0	2782080-48z	0	193162464-330z	0	709708608-576z
0	232944+36z	0	114572752+300z	0	1668859392+768z	0
2208-8z	0	20714904-78z	0	1432718208-240z	0	5268973584-336z
0	880992-24z	0	429089472-240z	0	6252791040-672z	0
3360+12z	0	39876480+192z	0	2761268220+840z	0	10150959840+1344z

The putative [96, 48, 20] doubly-even self-dual extremal code partitioned into two parts, one being the support of a minimum weight vector.

1	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	5480
0	0	0	0	116025	0
0	0	0	792900	0	3587400
0	0	1561845	0	38674095	0
0	708300	0	137608800	0	658256800
32505	0	149665635	0	3903490260	0
0	39125000	0	8007527400	0	38155220800
1067925	0	5189589375	0	134888439900	0
0	830347740	0	169334868960	0	807213470256
14160950	0	68596589625	0	1783592557125	0
0	6975586400	0	1423096463100	0	6783305787000
76613625	0	371194697250	0	9650968485075	0
0	24546567000	0	5007424075200	0	23868921318600
176560450	0	855437680590	0	22241422629840	0
0	37220804400	0	7593113876400	0	36193606924800

The putative [96, 48, 20] doubly-even self-dual extremal code partitioned into two parts, one being the support of a weight 24 vector. This distribution is not unique, but can be expressed in terms of a single variable.

1	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	2z
0	0	0	0	0	37260-12z	0
0	0	0	0	321264+30z	0	1145676+56z
0	0	0	1217160-40z	0	16655220-84z	0
0	0	1286712+30z	0	85790529	0	316577520-70z
0	346380-12z	0	157963540+140z	0	2309802600+504z	0
8280+2z	0	94051692-168z	0	6525055656-630z	0	23981575968-896z
0	13967532+84z	0	6854100372+140z	0	99844822152-168z	0
233430-16z	0	2401572600+294z	0	166178862045+1680z	0	611041252080+2758z
0	216634608-240z	0	105554190168-1400z	0	1538135889948-2940z	0
2119680+56z	0	22816881360+96z	0	1579342749840-630z	0	5806772474580-1400z
0	1288676016+336z	0	628553881536+2560z	0	9158856560676+6300z	0
8162217-112z	0	86140360656-1092z	0	5962012330338-3600z	0	21921113408544-5250z
0	3110887080-168z	0	1517024811960-1400z	0	22105251775440-3600z	0
12507216+140z	0	133641088392+1680z	0	9250021438128+6300z	0	34009965750656+9600z

REFERENCES

[1] E. ASSMUS AND V. PLESS, "On the covering radius of extremal self-dual codes", *IEEE Trans. Inform. Th.* 29 (1983),no. 3, 359–363.

[2] A.E. BROUWER AND T. VERHOEFF, "An updated table of minimum-distance bounds for binary linear codes", *IEEE Trans. Inform. Th.* 39 (1993) 662-677 as revised and updated at `http://www.win.tue.nl/math/dw/voorlincod.html`

[3] J.H. CONWAY AND V. PLESS, "An enumeration of self-dual codes", *IEEE Trans. Inform. Theory* IT-37 1222-1225 (1991)

[4] J.H. CONWAY, N.J.A. SLOANE *Sphere packings, Lattices and Groups*, Springer-Verlag, New York, 1988.

[5] W. C. HUFFMAN, "Automorphisms of Codes with Applications to Extremal Doubly Even Codes of Length 48", *IEEE Trans. Inform. Theory* IT-28 511-521

[6] H.V. KOCH, "Unimodular Lattices and Self-dual Codes", *Proceedings of the International Congress of Mathematicians* Berkeley, California, USA, 1986, pp 457-465.

[7] C.W.H. LAM, L. THIEL AND S. SWIERCZ, "The nonexistence of code words of weight 16 in a projective plane of order 10", J. Combin. Theory Ser. A 42 (1986) 207-214.

[8] C.W.H. LAM, L. THIEL, S. SWIERCZ AND J. MCKAY, "The nonexistence of ovals in a projective plane of order 10", Discrete Math. 45 (1983) 319-321.

[9] C.W.H. LAM, L. THIEL, S. SWIERCZ "The nonexistence of finite projective planes of order 10", Canad. J. Math. 41 (1989), no. 6, 1117–1123.

[10] V. PLESS *Introduction to the Theory of Error-Correcting Codes, second edition,* Wiley, New York, 1989.

[11] Juriaan Simonis, "MacWilliams Identities and Coordinate Partitions", *Linear Algebra Appl.* 216:81-91 (1995).

On Codes for Error Correction and Block Synchronization

Øyvind Ytrehus
University of Bergen, Department of Informatics,
Høyteknologisenteret, N-5020 Bergen, Norway.
Email: oyvind@ii.uib.no

Abstract *— A communication system or storage system can use a coset of a binary convolutional code for both symbol synchronization, block synchronization, and error control. To achieve block synchronization, no shift (shorter than a block length) of any coset sequence can itself be a coset sequence. In this work we provide bounds on the code rate of* self-synchronizing *cosets of convolutional codes, and give examples of such codes.*

I. Introduction

In this paper we study cosets of convolutional codes with self-synchronizing capabilities. Synchronization must be performed at different abstraction levels of a digital communication process.

- At the lowest level, an unsynchronized receiver needs to partition the signal waveforms received over a communication channel into chunks of time corresponding to symbol periods. This process, *symbol synchronization*, is briefly discussed below.
- At the next level, assuming that the information is coded in n-symbol blocks by a block code or a convolutional code, the task is to determine the boundaries between blocks. We refer to this as *block synchronization*. The study of block synchronization properties of (error-correcting) convolutional codes is the main topic of this paper.
- There is usually, depending on the application, a need for further synchronization at higher abstraction levels. For instance, the OSI reference model recommends a synchronization activity in the session layer. Such higher level synchronization is not considered in this paper.

Many communication systems and storage systems derive symbol synchronization from the transitions in the received symbol stream. To facilitate this symbol synchronization, it is often necessary to use a code which restricts the occurrences of certain ''bad'' sequences. Depending on the modulation scheme used, these ''bad'' sequences might be, for common examples, the all-zero sequence, or the all-zero *and* the all-one sequences. There are several ways to remove such sequences from the encoded data stream. Consider a system utilizing a linear binary error control code, i.e., a block code or a convolutional code. The code is unsuitable for symbol synchronization since it contains the all-zero sequence. However, the all-zero sequence (and other ''bad'' sequences) may be removed from the received symbol stream by using an appropriate coset of the code instead of the code itself. Other modulation schemes may require the removal of the all-zero sequence *and* the all-one sequence. This approach was studied in [1, 2, 3]. It was shown that cosets of codes in a particular class of convolutional codes, *ascetic convolutional codes* [4], are superior to other convolutional code cosets when it comes to restricting the lengths of zero-runs and one-runs in the encoded sequence. An (n, k) binary convolutional code is called *ascetic* if its set of branch labels does not contain all the 2^n binary vectors of length n. It can be seen that this definition is equivalent to stating that any (polynomial) parity check matrix of the code has at least one non-polynomial row. In this work, we study the block synchronization properties of convolutional code cosets.

II. Block Synchronization

Block synchronization can be achieved in many ways.

- Insertion of special synchronization sequences which do not occur elsewhere in the encoded sequence [5]. There is a trade-off between synchronization delay and the effective code rate.
- Use of a self-synchronizing code, such as for instance a *comma-free code* [6, Ch.3]. In general, such codes are not designed for error correction.
- Statistical synchronization: Use an error correcting code. Make a random guess on where the block boundaries are, and attempt to decode. If the channel appears to produce exceedingly many errors, try again.
- Use of a self-synchronizing error-correcting code. This is the approach that we will take.

Peterson and Weldon [7, Ch.12] discuss error correcting block codes capable of performing block synchronization. We observe that such codes exist, but the known constructions carry a substantial cost in terms of code rate reduction compared to non-synchronizing error correcting block codes.

Simple Example: Consider the convolutional code generated by

$$G(D) = (D \quad 1+D \quad 1+D \quad 1+D)$$

and with parity check matrix

$$\mathbf{H}(D) = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1+D & D & 0 & 0 \end{pmatrix}.$$

Figure 1 Trellis of simple example

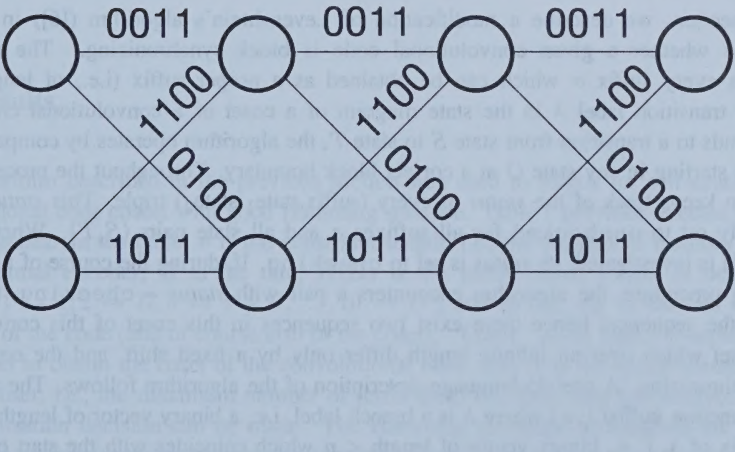


Fig. 1 shows the trellis of this simple example, where a coset is obtained by inverting the two last bits in every 4-bit branch label. We can verify from Fig. 1 that this coset has the following properties:

- The free Hamming distance is 7; largest possible among convolutional codes of rate 1/4 and with two decoder states.
- For symbol synchronization purposes, no sequence of consecutive zeros or consecutive ones has length larger than 4, and
- Block boundaries are uniquely determined after observation of at most 7 error free bits.

III. An Upper Bound

For self-synchronizing binary block codes of length n, it is known (see, for instance, [6]) that the number of codewords is upper bounded by

$$B(n) := (\sum_{d|n} \mu(d) 2^{n/d})/n,$$

the number of non-periodic cyclic equivalence classes (where $\mu(d)$ is the Möbius μ-function). We show that this bound also applies to cosets of (n, k, ν) convolutional codes regardless of constraint length ν, i.e., that also in this case $k \leq \log_2 B(n)$.

Theorem. *Let* C *be a binary* (n, k, ν) *convolutional code. A necessary, but not sufficient, condition for a coset of* C *to be self-synchronizing is that* $k \leq \log_2 B(n)$.

Proof. For each state S in the state diagram of $\mathcal{C}$, let $\mathcal{E}_S$ be the set of binary labels representing transitions from S back to S again, and let $\mathcal{E}$ be the union of all $\mathcal{E}_S$ when S runs through the state space. If $\mathcal{E}$ contains two vectors from the same non-periodic cyclic equivalence class, repeating these two vectors indefinitely yields two sequences which cannot be distinguished without knowledge of the block boundaries. We can verify that $\mathcal{E}$ is a linear block code of dimension k, and the theorem follows.

IV. A Generalization of Levenshtein's Algorithm

In this section, we describe a modification of Levenshtein's algorithm ([6]) in order to determine whether a given convolutional code is block synchronizing. The algorithm examines every suffix σ which can be obtained as a proper suffix (i.e., of length $< n$) of some transition label λ in the state diagram of a coset of a convolutional code. If λ corresponds to a transition from state S to state T, the algorithm operates by comparing σ to any path starting in any state Q at a correct block boundary. Throughout the procedure, the algorithm keeps track of the *status* of every (suffix,state$_1$,state$_2$) triple. This *status*$[\sigma, S, T]$ is initially set to `unchecked` for all suffixes σ and all state pairs (S, T). When a triple $([\sigma, S, T])$ is investigated, its *status* is set to `checking`. If, during the course of a recursive checking procedure, the algorithm encounters a pair with *status* = `checking`, there is a loop in the sequence; hence there exist two sequences in this coset of this convolutional code coset which over an infinite length differ only by a fixed shift, and the coset is not self-synchronizing. A pseudo-language description of the algorithm follows. The algorithm uses a function suffix(λ, σ) where λ is a branch label, i.e. a binary vector of length n, and σ is a prefix of λ, i. e., binary vector of length $< n$ which coincides with the start of λ. This function returns the suffix obtained by deleting the prefix σ from λ.

```
Algorithm Generalized_Levenshtein;
initialize: for each triple (suffix,state$_1$,state$_2$) do
                status[suffix,state$_1$,state$_2$]:=unchecked;
for each state $S$ in coset trellis do
   for each label $\lambda$ from state $S$ (to some state $T$) do
      for each proper suffix $\sigma$ of $\lambda$ do
         for each state $Q$ in coset trellis do
            FurtherCheck($\sigma, Q, T$);
End Algorithm Generalized_Levenshtein;
```

```
Algorithm FurtherCheck($\sigma$,ToState,NextToState);
if status[$\sigma$,ToState, NextToState]=unchecked then
   begin
   status[$\sigma$,ToState, NextToState]:=checking ;
   for each label $\lambda$ from state ToState (to some state NextNextToState) do
      if $\sigma$ is prefix of $\lambda$ then
         FurtherCheck(suffix($\lambda, \sigma$),NextToState,NextNextToState);
   status[$\sigma$,ToState,NextToState]:=OK;
   end
else
   if status[$\sigma$,ToState, NextToState]=checking then
      TERMINATE: code not self-synchronizing;
End Algorithm FurtherCheck;
```

It is possible to extend the algorithm so that it determines the maximum synchronization delay Δ, defined as the maximum number of bits that the algorithm needs to traverse in order to determine that a sequence is unique. For simplicity, the details are omitted.

V. Results

The algorithm described in the previous section was used to search for self-synchronizing convolutional code cosets with good Hamming distance. Table 1 provides selected examples of such codes. In the table, ν is the constraint length, i.e., the number of memory elements in a minimal encoder, R is the rate, $\mathbf{H}(\mathbf{D})$ is the parity check matrix in octal format $(0 \leftrightarrow 0, 1 \leftrightarrow 1, 2 \leftrightarrow D, 3 \leftrightarrow 1 + D, \ldots, 10 \leftrightarrow D^3$, and so on), d_H is the free Hamming distance of the code (and of course also of the coset), ''Coset'' is the binary n-tuple added to each label to obtain the coset of the convolutional code, and Δ is the synchronization delay of this coset, i.e., the maximum number of (error-free) bits that must be observed before a synchronization decision can be made. The Hamming distances in **boldface** are optimum for codes of these rates and constraint lengths. Thus in these cases the self-synchronizing property comes (roughly speaking) for free.

Table 1 A table of codes of selected rates

R	ν	$\mathbf{H(D)}$	Coset	d_H	Δ
1/3	2	$\begin{pmatrix} 3 & 1 & 2 \\ 0 & 3 & 1 \end{pmatrix}$	001	7	19
	3	$\begin{pmatrix} 2 & 3 & 1 \\ 1 & 4 & 6 \end{pmatrix}$	110	9	8
	4	$\begin{pmatrix} 7 & 1 & 5 \\ 0 & 7 & 5 \end{pmatrix}$	100	11	29
	5	$\begin{pmatrix} 7 & 1 & 6 \\ 3 & 16 & 13 \end{pmatrix}$	110	13	16
	6	$\begin{pmatrix} 17 & 2 & 11 \\ 0 & 11 & 13 \end{pmatrix}$	010	15	28
	7	$\begin{pmatrix} 7 & 6 & 7 \\ 2 & 73 & 33 \end{pmatrix}$	110	16	25
1/4	4	$\begin{pmatrix} 2 & 1 & 1 & 2 \\ 0 & 3 & 2 & 1 \\ 1 & 1 & 7 & 5 \end{pmatrix}$	0001	16	21
2/5	1	$\begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 & 1 \end{pmatrix}$	11110	4	8
	2	$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 3 & 1 & 1 & 2 \\ 0 & 0 & 2 & 3 & 1 \end{pmatrix}$	11000	6	10
3/6	0	$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$	111100	2	9

Table I (Continued) A table of codes of selected rates

R	ν	$\mathbf{H}(\mathbf{D})$	Coset	d_H	Δ
$3/6$	2	$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 2 & 0 & 3 & 0 \\ 0 & 0 & 3 & 3 & 1 & 1 \end{pmatrix}$	110000	4	11
$4/8$	1	$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 3 & 1 & 0 & 0 & 3 & 1 & 0 \end{pmatrix}$	00100000	4	20

Remarks to code searches

- Code rates: The bound in Section III allow the code rates 1/3, 2/5, 3/6, 4/7, 4/8, 5/9, and so on. From the code searches, we observe that for the low code rates (1/3, 1/4, 2/5) it seems to be easy to find self-synchronizing codes, often at no penalty in minimum distance. However, self-synchronizing codes of rate 3/6 are more scarce, and no codes of rate 4/7 were found for short constraint lengths. How rare are these codes? For instance, for the rate 1/3 parity check matrices considered, in the whole range of constraint lengths the proportion of cosets that are self-synchronizing was about one in four. In comparison, for rate 3/6, this proportion was on the order of 10^{-3} for constraint length 0, 10^{-4} for constraint length 1, and 10^{-5} for constraint length 2.
- For a given code there are often pairs of cosets which are self-synchronizing. For instance, if $\mathbf{1}$ is contained in the label code $\mathcal{E}$, but not in $\mathcal{E}_S$ for any state S (see Section III), then the trellis structure ensures that if a particular coset is self-synchronizing, then so is its complement. Other symmetries are also possible.
- Note that the complexity of the computer search is much higher for this problem than for the basic problem of finding codes with high minimum Hamming or Euclidean distance, since the order of the coordinates within a label is significant.
- Another empirical observation is that ascetic codes do not perform better than other codes with respect to block synchronization. Thus it appears in general to be hard to find a code which is simultaneously (close to) optimum with respect to minimum distance, symbol–, and block synchronization properties.

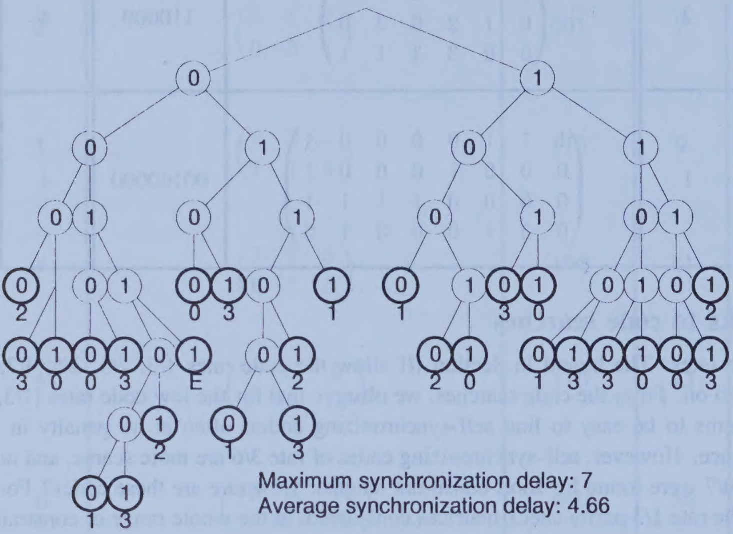
VI. Applications

This section discusses how the self-synchronizing codes can be used in a practical application.

Optimal synchronization: Synchronization trees

On a noiseless channel, the synchronization procedure can be performed by use of a synchronization tree. A synchronization tree for the simple example in Section II is shown in Fig. 2. By traversing the tree downwards from the top according to the received sequence, the block boundaries are determined when we reach a leaf node. The number below each leaf node shows the location of the block boundary relative to the leaf node. For instance, if we observe the sequence ...0111, the "1" below the associated leaf node denotes that the block boundary is located one bit back, i. e., after the subsequence ...011. Confer also with Fig. 1.

Figure 2 A synchronization tree



Maximum synchronization delay: 7
Average synchronization delay: 4.66

On a noisy channel, the synchronizer makes a hard decision on each channel symbol. If the raw error probability of this hard decision is p and the convolutional code coset requires at most Δ synchronizing symbols, the probability of wrong (single-attempt) synchronization is upper bounded by $1-(1-p)^{\Delta}$. In practical situations, it is natural to require that the channel is good enough for this upper bound to be small.

Anyway, we can recognize the synchronization delays achieved by the synchronization tree as a lower bound on the delay that is obtainable on a noisy channel.

The synchronization tree can be constructed by the algorithm in Section IV.

Statistical synchronization

Lorden *et. al.* [8] suggested to observe the syndrome sequence (obtained from hard-quantized data), in order to determine whether the sequence is correctly synchronized or not. This approach was extended by Mengali *et. al.* [9], who introduced soft-decision based synchronization.

In [9], a function $M(\mathrm{x})$ of the received sequence $\mathrm{x} = (x_1, \ldots, x_K)$ (of some pre-determined fixed length K) was compared with a threshold τ. The function $M(\mathrm{x})$ is defined as

$$M(\mathrm{x}) = \sum_{i=1}^{K} x_i(\bar{c}_i - \hat{c}_i), \tag{1}$$

where $\bar{c}_i$ is the symbol-by-symbol decision on the ith symbol and $\hat{c}_i$ is the ith symbol of the reencoded output of a Viterbi decoder. The receiver is assumed to be synchronized if and only if $M(\mathrm{x}) < \tau$.

In [8] and [9], the *short average run length (SARL)* and the *long average run length (LARL)* were used as figures of merit for synchronization schemes. The SARL is the average number of bits that must be observed before synchronization is established, while the LARL is the average number of bits that are transmitted before the synchronizer erroneously, due to the impact of noise on $M(\mathrm{x})$ (see equation (1)), declares an out-of-synch error. Of course we want the SARL to be small and the LARL to be large. If the receiver is synchronized, it does not matter to equation (1) whether the code is self-synchronizing or not. However, if the receiver is not synchronized, we would expect $M(\mathrm{x})$ to grow faster with K for self-synchronizing codes than for other codes. Thus, for a self-synchronizing code, the threshold τ can be adjusted to improve the LARL/SARL ratio.

It is a future project to verify this heuristic argument by computer simulation.

Relation to CDMA sequences

This work is distantly related to the problem of finding CDMA sequences. A family of CDMA sequences of length N can be regarded as a code with Hamming distance d between any (code word, shift of a code word) pair. In the CDMA terminology, the family of sequences should have low auto— and cross correlation $\theta = N - 2d$ where $\sqrt{N} \leq \theta \leq 2\sqrt{N}$ is the most interesting range of θ [10].

References

[1] Kjell Jørgen Hole and Øyvind Ytrehus. Improved coding techniques for precoded partial-response channels. *IEEE Trans. on Information Theory*, IT-40:482–493, March 1994.

[2] Kjell Jørgen Hole. Cosets of convolutional codes with short maximum zero-run lengths. *IEEE Trans. on Information Theory*, IT-41:1145–1150, July 1995.

[3] Kjell Jørgen Hole and Øyvind Ytrehus. Cosets of convolutional codes with least possible maximum zero- and one-run lengths. *IEEE Trans. on Information Theory*, 1998. To appear.

[4] Øyvind Ytrehus. Ascetic convolutional codes. In *Proc. 33rd Allerton Conference on Communication Control and Coding, Monticello, Illinois, USA, Oct. 4-7*, 1995.

[5] J. L. Massey. Optimum frame synchronization. *IEEE Trans. on Communications*, COM-20:115–119, April 1972.

[6] S. W. Golomb, R. E. Peile, and R. A. Scholtz. *Basic Concepts in Information Theory and Coding*. Plenum, 1994.

[7] W. W. Peterson and E. J. Weldon. *Error-Correcting Codes, 2nd. Ed.* MIT Press, 1972.

[8] G. Lorden, R. J. McEliece, and L. Swanson. Node synchronization for the Viterbi decoder. *IEEE Trans. on Communications*, COM-32:524–531, May 1984.

[9] U. Mengali, R. Pellizzoni, and A. Spalvieri. Soft-decision-based node synchronization for Viterbi decoders. *IEEE Trans. on Communications*, 43:2532–2539, September 1995.

[10] T. Helleseth and P. V. Kumar. Pseudonoise sequences. In J. D. Gibson, editor, *The Mobile Communications Handbook*, chapter 8, pages 110–122. CRC/IEEE, 1996.

Combined Array Signal Processing and Space-Time Coding

V. Tarokh, A. Naguib, N. Seshadri and A.R. Calderbank
AT&T Labs-Research
180 Park Avenue
Florham Park, NJ 07932.

Abstract

Information theory provides measures of capacity for wireless communication systems employing multiple transmit and receive antennas, and channel capacity is high when there are many antennas at both the transmitter and receiver [4, 12]. The standard approach to increasing data rate on wireless channels has been array processing at the receiver, and less attention has been paid to coding techniques appropriate to multiple transmit antennas. Space-time codes [10] introduce temporal and spatial correlation into signals transmitted from different antennas, so as to provide diversity at the receiver, and coding gain over an uncoded system. These codes have been shown to achieve substantial improvements in the data rate and/or reliability of narrowband wireless channels employing small numbers (1-4) of transmit and receive antennas. For large number of transmit antennas and at high bandwidth efficiencies, the receiver becomes too complex whenever correlation across transmit antennas is introduced.

This paper dramatically reduces encoding and decoding complexity by partitioning antennas at the transmitter into small groups, and using individual space-time codes, called component codes, to transmit information from each group of antennas. At the receiver, an individual space-time code is decoded by a novel linear processing technique that suppresses signals transmitted by other groups of antennas by treating them as interference. A simple receiver structure is derived that provides diversity and coding gain over uncoded systems. This combination of array signal processing at the receiver and coding techniques for multiple transmit antennas provides reliable and very high data rate communication over narrowband wireless channels. A refinement of this basic structure gives rise to a multi-layered space-time architecture that both generalizes and improves upon the layered space-time architecture proposed by Foschini [3].

1 Introduction.

Physical limitations on wireless channels present a fundamental technical challenge to reliable communication. Bandwidth limitations, propagation loss, time variance, noise, interference, and multipath fading make the wireless channel a narrow pipe that does not easily accommodate the flow of data. Further challenges come from power limitation and size and speed of devices in wireless portables.

Current cellular standards such as IS-136 support circuit data and fax services at the rate 9.6 kpbs, and a packet data mode is being standardized. Recently, there has been growing interest in providing a broad range of services, including high quality voice and wireless data rates of 64-128 kbps using the cellular (850-MHz) and PCS (1.9 GHz) spectra. Within IS-136 standard, the approach taken to improve voice quality in [8], was to expand the signal constellation size and use powerful coding and modulation techniques. However, there is a substantial SNR penalty to be paid for increasing constellation size, and this approach is not capable of achieving higher speed data services in the range of 144 kbps.

Deploying multiple transmit antennas at both the base and remote stations increases the capacity of wireless channels [4, 7, 12], and the standard approach to exploit this capacity is linear processing at the receiver. Transmit diversity has been explored by Wittneben [15], and and later independently by Seshadri and Winters[9] as a special case. These papers approached transmit diversity from a signal processing point of view. Space-time codes [10] combine signal processing at the receiver with coding techniques appropriate to multiple transmit antennas, and this approach provides significant gain over [9, 15]. Specific space-time codes designed for 2-4 transmit antennas perform extremely well in slowly varying fading environments (such as indoor transmission) and come with in 2-3 dB of the outage capacity as computed by Telatar [12] and independently by Foschini and Gans [4]. The bandwidth efficiency of these codes are about 3-4 times that of the current systems. The most important contributor to this performance is diversity, which we may think of as providing some less attenuated replica of the transmitted signal to the receiver. The space-time codes presented in [10] provide the best possible trade-off between constellation size, data rate, diversity advantage, and trellis complexity.

When the number of transmit antennas is fixed, decoding complexity (measured by the number of trellis states in the decoder) increases exponentially with transmission rate. This can be finessed to some extent by designing space-time codes with a multi-level structure and adopting multi-stage decoding as described in [10]. For a moderate number of transmit antennas (3-6), this method provides higher data rates while reducing the decoding complexity. However, there is a penalty to be paid for simplified decoding. Multi-stage decoding is suboptimal, due in part to magnification in the error coefficient, and this performance penalty means that alternative solutions are needed for achieving very high data rates.

In order to achieve very high data rates on narrowband wireless channels, we need many antennas at both transmitter and receiver. Consider a wireless communication system employing n transmit and m receive antennas, where the subchannel between each transmit and receive antenna is quasi-static Rayleigh, flat and independent of others. If n is fixed, then the capacity increases only logarithmically with m [4, 12]. On the other hand, if m is fixed then intuition suggests that, there must come a point where adding more transmit antennas will not make much of difference and this can be seen in the mathematics of outage capacity. For instance if $m = 1$, then Foschini and Gans [4] prove that the capacity of the aforementioned system is a random variable of the form $\log_2(1 + \frac{\chi^2_{2n}}{2n}\mathrm{SNR})$, where χ^2_{2n} is a random variable formed by summing the squares of $2n$ independent Gaussian random variables with mean zero and variance one. This means that by the strong law of large number $\frac{\chi^2_{2n}}{2n} \to 1$ in distribution. Practically speaking, for $n = 4$, $\frac{\chi^2_{2n}}{2n} \simeq 1$ and the capacity is the familiar Gaussian capacity $\log_2(1 + \mathrm{SNR})$ per complex dimension. Thus, in the presence of one receive antenna, little can be gained in terms of outage capacity by using more than 4 transmit antennas. A similar argument shows that if there are two receive antennas, almost all the capacity increase can be obtained using $n = 6$ transmit antennas. These observations also follow from the capacity plots given by Telatar[12].

If n increases and $m \geq n$, then information theory [4, 12] shows that the capacity of the system increases at least linearly as a function of n. Thus it makes sense to increase the number of antennas at both receiver and transmitter in order to obtain higher capacities. Many antennas at both transmitter and receiver creates a multiple-input multiple-output system where the number of degrees of freedom is given by the product of the number of transmit and receive antennas. The approach pursued here is to employ a small fraction of these degrees of freedom for improving the performance while using the remaining degrees of freedom to achieve higher data rates. This leads to a combination of space-time coding at the transmitter and array processing at the receiver which achieves high data rates and operates close to the limits given by outage capacity.

Foschini considered this problem in [3]. He proposed a multi-layered structure that in principal can achieve a tight lower bound on the capacity. If n transmit and n receive antennas are used, then at the receiver the transmitted signal from transmit antenna one is treated as the desired signal, while the signals transmitted from other transmit antennas are treated as interference. Linear processing is then used to suppress the interfering signals using n receive antennas, providing a diversity advantage of one. Once the signal transmitted from antenna one is correctly detected, the signal transmitted from antenna 2 is treated as the desired signal while the signals transmitted from transmit antennas $3, \cdots, n$ are treated as interference. Linear processing is then applied to suppress the interfering signals using n receive antennas. This provides a diversity advantage of two. This process is repeated until all the transmitted signals are detected. Clearly, the dominant diversity in this architecture is one. In this light, very long frames of data combined with powerful coding techniques are needed to achieve the lower bound on outage capacity [3].

The approach described above is purely a signal processing one. In this work, we depart from a purely signal processing point of view and approach the problem from an array signal processing combined with channel coding perspective. Antennas at the transmitter are partitioned into small groups, and individual space-time codes, called the *component codes*, are used to transmit information from each group of antennas. At the receiver an individual space-time code is decoded by a novel linear array processing technique, called the *group interference suppression method*, that suppresses signals transmitted by other groups of antennas by treating them as interference. This method involves only elementary linear algebra, and is similar in spirit to techniques developed for multi-user detection [6, 13] and to temporal and spatial techniques developed for interference suppression [1, 14]. A simple receiver structure is derived that provides diversity and coding gain over uncoded systems with a given diversity advantage. This combination of array processing at the receiver and coding techniques for multiple transmit antennas provides reliable and very high data rate communication over wireless channels. One obvious advantage of the group interference suppression method over the architecture of Foschini [3] is that the number of receive antennas can be less than the number of transmit antennas.

2 The Channel Model

We model a wireless communication system with n antennas at the base and m antennas at the mobile station. Data is encoded using a channel code. The encoded data goes through a serial to parallel converter and is divided into n streams of data. Each stream of data is used as the input to a pulse shaper. The output of each shaper is then modulated. At each time slot t the output of modulator i is a signal c_t^i that is transmitted using transmit antenna (Tx antenna) i for $1 \leq i \leq n$. We emphasize that the n signals are transmitted simultaneously each from a different transmit antenna and that all these signals have the same transmission period T. The signal at each receive antenna is a noisy superposition of the n transmitted signals corrupted by Rayleigh or Rician fading.

At the receiver, the demodulator computes a decision statistic based on the signals received at each receive antenna $1 \leq j \leq m$. The signal r_t^j received by antenna j at time t is given by

$$r_t^j = \sum_{i=1}^{n} \alpha_{i,j} c_t^i + \eta_t^j. \tag{1}$$

The coefficient $\alpha_{i,j}$ is the path gain from transmit antenna i to receive antenna j and η_t^j is the noise for the channel between transmit antennas and receive antenna j at time t. The path gains $\alpha_{i,j}$ are modeled as samples of independent complex Gaussian random variables with mean zero and variance 0.5 per dimension. This is equivalent to the assumption that signals transmitted from different antennas undergo independent fades. The noise quantities $\eta_t^j, j = 1, \cdots, m, t = 1, 2, \cdots, l$ are samples of independent complex Gaussian random variable with mean zero and variance $N_0/2$ per dimension. We further assume that $\alpha_{i,j}$ are constant during a frame and vary from one frame to another (quasi-static fading).

3 Combined Array Processing and Space-Time Coding

A *space-time product encoder* accepts a block of B input bits in each time slot t and these bits are divided into q strings of lengths $B_1, B_2, \cdots, B_q$ with $B_1 + B_2 + \cdots + B_q = B$. Consider the communication model given Section 2, with n antennas at the base station divided into q groups comprising $n_1, n_2, \cdots, n_q$ antennas with $n_1 + n_2 + \cdots + n_q = n$. Each block B_j, $1 \leq j \leq q$ is then encoded by a space-time encoder $\mathcal{C}_j$. The output of $\mathcal{C}_j$ goes through a serial to parallel converter and provides n_j sequences of constellation symbols for $1 \leq j \leq q$. This gives a total of n sequences of constellation symbols that are transmitted simultaneously from antennas $1, 2, \cdots, n$. We view a space-time product encoder as a set of q space-time encoders called the *component codes* operating in parallel on the same wireless communication channel, with each encoder using n_j transmit and m receive antennas for $1 \leq j \leq q$. It will be denoted by $\mathcal{C}_1 \times \cdots \times \mathcal{C}_q$.

The partition of transmit antennas into small groups, and the use of component space-time codes to transmit information from each group of antennas fits naturally in the framework of concatenated or product codes [2]. Product codes are applied for interference suppression in [1], where Calderbank, Pottie and Seshadri use the product code construction of first order Reed-Muller codes for a combination of coding gain and interference suppression.

The task is to recover the transmitted data at the receiver. One approach is to jointly decode for the transmitted codeword, but the difficulty here is decoding complexity. Indeed if we require a diversity of rm, where $r \leq \min_j(n_j)$, then the complexity of the trellis of $\mathcal{C}_j$ is at least $2^{B_j(r-1)}$ states [10] and the complexity of the product code is at least $2^{B(r-1)}$ states. This means that if B is very large, the scheme is too complex to implement.

The approach, we advocate is *group interference suppression*. The idea is to decode each code $\mathcal{C}_j$ separately while treating signals from other component codes as interference. Without loss of generality, we take $j = 1$ and look to decode $\mathcal{C}_1$. There are $n - n_1$ interfering codewords. We assume that there are $m \geq n - n_1 + 1$ receive antennas, and form $m - n + n_1$ groups $G_1, G_2, \cdots, G_{m-n+n_1}$ of receive antennas. The group G_p contains the first $n - n_1$ receive antennas, labelled $1, 2, \cdots, n - n_1$, together with receive antenna $n - n_1 + p$.

It is sufficient to demonstrate that the antennas in group G_1 can be used to suppress codewords from component codes $\mathcal{C}_j$, $j \neq 1$. The signal r_t^j received by antenna j at time t is given by

$$r_t^j = \sum_{i=1}^{n_1} \alpha_{i,j} c_t^i + \sum_{i=n_1+1}^{n} \alpha_{i,j} c_t^i + \eta_t^j, \tag{2}$$

where c_t^i is the signal transmitted by antenna i at time t and $\alpha_{i,j}$ is the path gain from transmit antenna i to receive antenna j. We assume that the receiver knows the path gains $\alpha_{i,j}$ (perfect CSI) and describe how to suppress interference from component codes other than $\mathcal{C}_1$.

To this end, we observe that the rank of the matrix

$$\Lambda(\mathcal{C}_1, G_1) = \begin{pmatrix} \alpha_{n_1+1,1} & \alpha_{n_1+2,1} & \cdots & \cdots & \alpha_{n,1} \\ \alpha_{n_1+1,2} & \alpha_{n_1+2,2} & \cdots & \cdots & \alpha_{n,2} \\ \alpha_{n_1+1,3} & \alpha_{n_1+2,3} & \cdots & \cdots & \alpha_{n,3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \alpha_{n_1+1,n-n_1+1} & \alpha_{n_1+2,n-n_1+1} & \cdots & \cdots & \alpha_{n,n-n_1+1} \end{pmatrix}, \tag{3}$$

is less than or equal to $n - n_1$. Indeed, the rank of $\Lambda(\mathcal{C}_1, G_1)$ is equal to $n - n_1$ with probability one. Thus, one can find a vector

$$\mathbf{w}_1 = (w_1^1, w_2^1, \cdots, w_{n-n_1+1}^1),$$

such that $||\mathrm{w}_1|| = 1$ and

$$\mathrm{w}_1 \cdot \Lambda(\mathcal{C}_1, G_1) = (0, 0, \cdots, 0). \tag{4}$$

The vector w_1 is (with probability one) unique up to multiplication by $exp(\mathrm{j}\theta_1)$, where $\mathrm{j} = \sqrt{-1}$. The phase ambiguity θ_1 will not affect the decoding metric derived below and in fact we may choose an arbitrary value for θ_1, when specifying w_1.

We observe from (4) that

$$w_1^1 \alpha_{j,1} + w_2^1 \alpha_{j,2} + \cdots + w^1_{n-n_1+1} \alpha_{j,n-n_1+1} = 0.$$

for $j = n_1 + 1, \cdots, n$. Now we use w_1 to combine the received signals $r_t^j, j = 1, 2, \cdots, n - n_1 + 1$. Let

$$\beta_{i,1} = w_1^1 \alpha_{i,1} + w_2^1 \alpha_{i,2} + \cdots + w^1_{n-n_1+1} \alpha_{i,n-n_1+1}, \tag{5}$$

for $i = 1, \cdots, n_1$. and define

$$\mathrm{r}_t(\mathcal{C}_1, G_1) = w_1^1 r_t^1 + w_2^1 r_t^2 + \cdots + w^1_{n-n_1+1} r_t^{n-n_1+1}, \tag{6}$$

then it follows easily that

$$\mathrm{r}_t(\mathcal{C}_1, G_1) = \beta_{1,1} c_t^1 + \beta_{2,1} c_t^2 + \cdots + \beta_{n_1,1} c_t^{n_1} + \eta_t(\mathcal{C}_1, G_1), \tag{7}$$

where

$$\eta_t(\mathcal{C}_1, G_1) = w_1^1 \eta_t^1 + w_2^1 \eta_t^2 + \cdots + w^1_{n-n_1+1} \eta_t^{n-n_1+1} \tag{8}$$

This is an equation where signals from component codes other than $\mathcal{C}_1$ are suppressed. Furthermore, in this suppression, only the group of receive antennas corresponding to G_1 is used. By using the group of antennas corresponding to $G_2, G_3, \cdots, G_{m-n+n_1}$, we arrive at $m - n + n_1$ equations similar to that given in (7). Indeed, by choosing

$$\mathrm{w}_p = (w_1^p, w_2^p, \cdots, w^p_{n-n_1}, w^p_{n-n_1+p})$$

with $||\mathrm{w}_p|| = 1$, in a manner similar to w_1, we have

$$\mathrm{r}_t(\mathcal{C}_1, G_p) = \beta_{1,p} c_t^1 + \beta_{2,p} c_t^2 + \cdots + \beta_{n_1,p} c_t^{n_1} + \eta_t(\mathcal{C}_1, G_p), \tag{9}$$

where

$$\beta_{k,p} = w_1^p \alpha_{k,1} + w_2^p \alpha_{k,2} + \cdots + w^p_{n-n_1} \alpha_{k,n-n_1} + w^p_{n-n_1+p} \alpha_{k,n-n_1+p}, \tag{10}$$

and

$$\eta_t(\mathcal{C}_1, G_p) = w_1^p \eta_t^1 + w_2^p \eta_t^2 + \cdots + w^p_{n-n_1+p} \eta_t^{n-n_1+p} \tag{11}$$

for $k = 1, 2, \cdots, n_1$ and $p = 1, 2, \cdots, m - n + n_1$.

Assuming perfect CSI, the path gains $\alpha_{i,j}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, m$ are all known. The vectors

$$\mathrm{w}_p, \quad p = 1, 2, \cdots, m - n + n_1,$$

the constants

$$\beta_{i,j}, \quad i = 1, 2, \cdots, n_1, \; j = 1, 2, \cdots, m - n + n_1,$$

and the estimates

$$\mathrm{r}_t(\mathcal{C}_1, G_1), \mathrm{r}_t(\mathcal{C}_1, G_2), \cdots, \mathrm{r}_t(\mathcal{C}_1, G_{m-n+n_1})$$

are easily computed. Only codewords from $\mathcal{C}_1$ contribute to the estimates

$$\mathrm{r}_t(\mathcal{C}_1, G_p), \; p = 1, 2, \cdots, m - n + n_1.$$

We further notice that the noise terms $\eta_t(\mathcal{C}_1, G_p), p = 1, 2, \cdots, m - n + n_1$ are not independent. However, the vectors w_p, $p = 1, 2, \cdots, n_1 + m - n$ are known, so we have complete information about the joint statistics of $\eta_t(\mathcal{C}_1, G_p), p = 1, 2, \cdots, m - n + n_1$ and the derivation of MLSE receiver for $\mathcal{C}_1$ is straightforward (for more details see [11]).

4 The Receiver Structure

Let $N_0 D$ denote the covariance matrix of the noise terms $\eta_t(\mathcal{C}_1, G_p)$, for $p = 1, 2, \cdots, m-n+n_1$. The i-th diagonal element $D_{i,i} = ||\mathrm{w}_i||^2 = 1$, and for $i \neq j$, we have

$$D_{i,j} = \sum_{r=1}^{n-n_1} w_r^i \bar{w}_r^j.$$

Let $D^{-1} = F \cdot F^*$ where F is an upper-triangular matrix, and F^* denotes the Hermitian of F. Let

$$\mathbf{R}_t = (\mathrm{r}_t(\mathcal{C}_1, G_1), \mathrm{r}_t(\mathcal{C}_1, G_2), \cdots, \mathrm{r}_t(\mathcal{C}_1, G_{n_1}))$$

Suppose that $c_t^1, \cdots, c_t^{n_1}$ were transmitted from antennas $1, 2, \cdots, n_1$ respectively. Since $\beta_{i,j}$, $i = 1, 2, \cdots, n_1$, $j = 1, 2, \cdots, m-n+n_1$ are known, the random vector

$$\mathbf{R}_t = (\mathrm{r}_t(\mathcal{C}_1, G_1), \mathrm{r}_t(\mathcal{C}_1, G_2), \cdots, \mathrm{r}_t(\mathcal{C}_1, G_{n_1}))$$

has mean

$$\mu_t = \left(\sum_{i=1}^{n_1} \beta_{i,1} c_t^i, \sum_{i=1}^{n_1} \beta_{i,2} c_t^i, \cdots, \sum_{i=1}^{n_1} \beta_{i,m-n+n_1} c_t^i \right)$$

and variance 0.5 per dimension. We may calculate the random vector $(\mathbf{R}_t - \mu_t)F$ and this vector has mean zero and variance $N_0/2$ per dimension. Moreover, the components of $(\mathbf{R}_t - \mu_t)F$ are i.i.d complex Gaussian random variables.

The optimum detection rule is then to compute

$$\sum_{t=1}^{l} |(\mathbf{R}_t - \mu_t)F|^2,$$

for each codeword

$$c_1^1 c_1^2 \cdots c_1^{n_1} c_2^1 c_2^2 \cdots c_2^{n_1} \cdots c_l^1 c_l^2 \cdots c_l^{n_1},$$

and decide in favor of the codeword that minimizes this sum. Thus the optimum branch metric for Viterbi decoding is

$$|(\mathbf{R}_t - \mu_t)F|^2 = |\mathbf{R}_t F - (c_t^1, c_t^2, \cdots, c_t^{n_1})\Omega F|^2,$$

where Ω is the $n_1 \times (m-n+n_1)$ matrix

$$\Omega = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \cdots & \beta_{1,m-n+n_1} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \cdots & \beta_{2,m-n+n_1} \\ \beta_{3,1} & \beta_{3,2} & \cdots & \cdots & \beta_{3,m-n+n_1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \beta_{n_1,1} & \beta_{n_1,2} & \cdots & \cdots & \beta_{n_1,m-n+n_1} \end{pmatrix}. \qquad (12)$$

It can be seen that the above metric is independent of $\theta_1, \theta_2, \cdots, \theta_{m-n+n_1}$, where θ_p is the phase ambiguity in the definition of w_p.

The space-time code C_1 is described by a trellis, and decoding is by applying the Viterbi algorithm using the above metric. For each frame, the matrices F and ΩF are precomputed while the vector $R_t F$ has to be computed once per time slot t. These are straightforward computations that can be easily implemented both in hardware and software.

5 Performance Analysis

Having derived the optimum MLSE receiver for each component code, we now analyze the error performance of the system. The next result expresses the fact that the group interference suppression method does indeed suppress the interference from the other component codes $C_2, \cdots, C_{n_1}$ at no expense to the performance of the component code C_1. The proof is given in [11].

Theorem 5.1 *Let $E_i, 1 \leq i \leq n$, denote the average transmit power out of antenna i. The probability of error P_1 for C_1 using combined array processing and space-time coding equals the probability of error P_2 for C_1 when used alone for communications using n_1 transmit and $m - n + n_1$ receive antennas where $E_i, 1 \leq i \leq n_1$, is the average transmit power out of antenna i.*

Theorem 5.1 and the theory of group interference suppression reduces code design for a multiple antenna communication system with n transmit and m receive antennas to that of designing codes for communication systems with n_i transmit and $n_i + m - n$, $i = 1, 2, \cdots, q$, receive antennas where $\sum_{i=1}^{q} n_i = n$ and $n_i \geq n - m + 1$. Using this insight, we may design a multi-layered space-time coded modulation scheme.

Suppose that C_1 is decoded correctly using combined array processing and space-time coding. By Theorem 5.1, the space-time code C_1 affords a diversity gain of $n_1 \times (n_1 + m - n)$. After decoding C_1, we may subtract the contribution of these codewords to signals received at different antennas. This gives a communication system with $n - n_1$ transmit and m receive antennas. We next use combined array processing and space-time coding to decode C_2. The space-time code C_2 affords a diversity gain of $n_2 \times (n_2 + n_1 + m - n)$. Proceeding in this manner, we observe that by subtracting the contribution of previously decoded code streams $C_j, j \leq k - 1$ to the received signals at different antennas, the space-time code C_k affords a diversity gain of $n_k \times (n_1 + \cdots + n_k + m - n)$.

We can choose space-time codes C_i, $1 \leq i \leq q$ to provide these diversity gains and such that the sequence

$$n_1(n_1 + m - n), n_2(n_2 + n_1 + m - n), \cdots, n_k(n_1 + \cdots + n_k + m - n), \cdots, n_q m$$

be an increasing sequence. Assuming there was no decoding error in steps $1, 2, \cdots, k-1$, then at decoding step k, the probability of error for the component code C_k is by Theorem 5.1 equal to the probability of error for C_k when employed in a communication system using n_k transmit and $(n_1 + \cdots + n_k + m - n)$ receive antennas.

Since the diversity in each decoding stage k is more than that of the previous decoding stage $k - 1$, the transmit power out of each antennas at level k can be substantially less than that of previous layer. Thus the transmitter should divide the available transmit power amongst different antennas in an unequal manner. Power allocation for this scenario is straightforward. In fact powers at different levels could be allocated based on the diversity gains. In this way, the allocated powers may decrease logarithmically in terms of the diversity gains. Other approaches are also possible.

We illustrate the construction of multi-layered space-time coded modulation schemes by an example. Other examples are given in [11].

Example: Here, 4 transmit and 4 receive antennas are used. The transmission rate is 4 bits/sec/Hz. Let C denote the code given in Figure 1. We use the product code $C_1 \times C_2$ where $C_1 = C_2 = C$ for transmission of 4 bits/sec/Hz. At each time slot, upon the arrival of the 4 bits of the input data, the first two bits are used as the input to the encoder of C_1 and the encoded symbols are transmitted by antennas one and two. The second two bits are used by as the input to the encoder of C_2 and the encoded signals are transmitted by antennas 3 and 4. We assume that the average powers radiated from antennas 1 and 2 are equal but each is twice as much as the average power radiated from antennas 3 and 4. At the decoder, group interference suppression is used to suppress $\mathcal{C}_2$ and decode $\mathcal{C}_1$. Upon decoding $\mathcal{C}_1$, the contributions of the codewords transmitted from antennas one and two are subtracted from the received signals. Finally, $\mathcal{C}_2$ is decoded. In Figure 2, we provide simulation results to demonstrate the performance of this multi-layered space-time coded architecture. The frame structure is the same as that of IS-136. For comparison, we provide the graph of the outage capacity versus the signal to noise ratio for 4 transmit and 4 receive antennas as computed by Foschini and Gans [4]. We observe that for a frame error probability of 10^{-1}, we are about 6 dB away from the capacity.

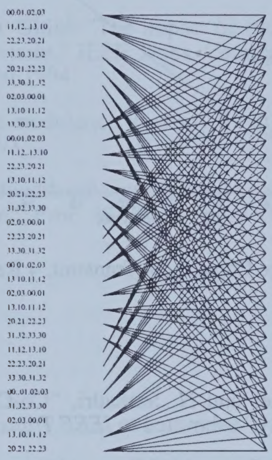

00.01.02.03
11.12..13.10
22.23.20.21
33.30.31.32
20.21.22.23
33.30.31.32
02.03.00.01
13.10.11.12
33.30.31.32
00.01.02.03
11.12..13.10
22.23.20.21
13.10.11.12
20.21.22.23
31.32.33.30
02.03.00.01
22.23.20.21
33.30.31.32
00.01.02.03
13.10.11.12
02.03.00.01
13.10.11.12
20.21.22.23
31.32.33.30
11.12.13.10
22.23.20.21
33.30.31.32
00..01.02.03
31.32.33.30
02.03.00.01
13.10.11.12
20.21.22.23

Figure 1: 4-PSK, 32-State Code

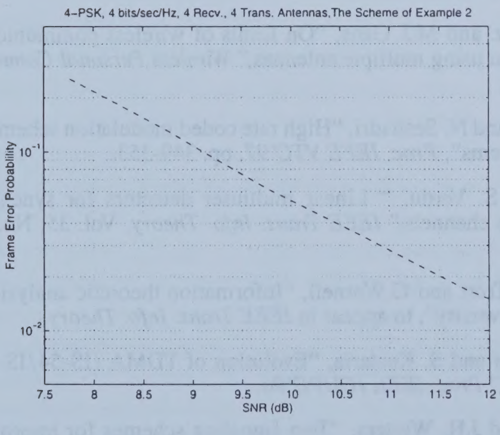


Figure 2: The Performance of The Scheme of Example 2

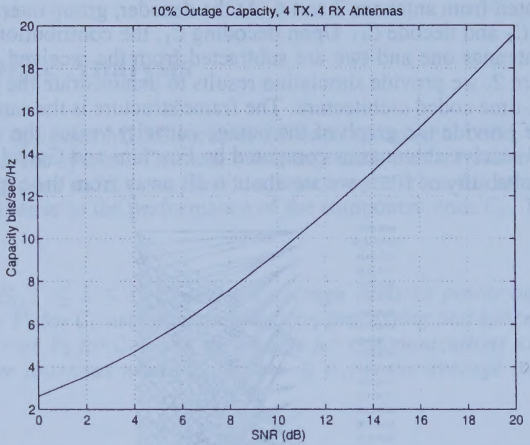


Figure 3: Outage Capacity, 4 Transmit, 4 Receive Antennas

References

[1] A.R. Calderbank, G. Pottie, and N. Seshadri, "Co-Channel interference suppression through time/space diversity," submitted to *IEEE Trans. Info. Theory*.

[2] G.D. Forney, *Concatenated Codes*, MIT Press, 1966.

[3] G.J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Technical Journal*, Vol 1., No. 2, Autumn 1996.

[4] G.J. Foschini, Jr. and M.J. Gans, "On limits of wireless communication in a fading environment when using multiple antennas," *Wireless Personal Communication*, submitted for publication.

[5] E.A. Gelblum and N. Seshadri, "High rate coded modulation schemes for 16 kbps speech in wireless systems", *Proc. IEEE VTC'97*, pp. 349-353.

[6] R. Lupas and S. Verdu, " Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Info. Theory*, Vol. 35, No. 1, pp. 123-136, Jan. 1989.

[7] A. Narula, M.Trott and G.Wornell, "Information theoretic analysis of multiple-antenna transmission diversity", to appear in *IEEE Trans. Info. Theory*.

[8] N. Sollenberger and S. Kustaria, "Evolution of TDMA (IS-54/IS-136) to foster further growth of PCS," *Proc. IEEE ICUPC'96*.

[9] N. Seshadri and J.H. Winters, "Two signaling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity", *International Journal of Wireless Information Networks,* Vol. 1, No. 1, 1994.

[10] V. Tarokh, N. Seshadri and A.R. Calderbank, "Space-Time codes for high data rate wireless communication: Performance analysis and code construction," to appear in *IEEE Trans. Info. Theory*.

[11] V. Tarokh, A. Naguib, N. Seshadri and A.R. Calderbank, "Space-Time codes for high data rate wireless communication: Combined array processing and space-time coding" submitted to *IEEE Trans. Info. Theory.*

[12] E. Telatar, "Capacity of multi-antenna Gaussian channels", AT&T-Bell Laboratories Internal Technical Memorandum, June 1995.

[13] M.K. Varanasi, "Group detection for synchronous Gaussian code-division multiple-access channels," *IEEE Trans. Info. Theory*, Vol. 41, No. 4, pp. 1083-1096, July 1995.

[14] J. Winters, J. Salz and R.D. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," IEEE Trans. Communications, Vol 42. No. 2/3/4, pp. 1740-1751, Feb/March/April 1994.

[15] A. Wittneben, "Base station modulation diversity for digital SIMULCAST," *Proc. IEEE' VTC*, pp. 505-511, May 1993.

[16] A. Wittneben , "A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation", *Proc. IEEE'ICC*, pp. 1630-1634, 1993.

Adaptive Multiuser Detection in Non-Gaussian Channels [*]

Xiaodong Wang and *H. Vincent Poor*
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544

Abstract

Robust multiuser detection techniques are proposed for combating multiple-access interference and impulsive ambient noise in CDMA communication systems. These techniques are based on the method of M-estimation for robust regression. Analytical and simulation results show that the proposed robust techniques offer significant performance gain over linear multiuser detectors in impulsive noise, with little attendant increase in computational complexity. The robust multiuser detection techniques can be implemented *blind-adaptively* by using a subspace technique.

1 Introduction

The study of multiuser detection has primarily focused on the situation in which the ambient noise is additive white Gaussian noise (AWGN). However, in many wireless systems where multiuser detection techniques may be applied, the ambient channel noise is known through experimental measurements to be decidedly non-Gaussian, due largely to impulsive phenomena [4, 5]. It is widely known in the single-user context that non-Gaussian noise can be quite detrimental to the performance of conventional systems based on the Gaussian assumption, whereas it can actually be beneficial to performance if appropriately modeled and ameliorated.

In view of the lack of realism of an AWGN model for ambient noise arising in many practical channels in which multiuser detection techniques may be applied, natural questions arise concerning the applicability, robustness and performance of multiuser detection techniques for non-Gaussian multiple-access channels. Impulsive noise can severely degrade the error probability for a given level of ambient noise variance. In the context of multiple-access capability, this implies that fewer users can be supported with conventional detection in an impulsive channel than in a Gaussian channel. However, since non-Gaussian noise can, in fact, be beneficial to system performance if properly treated, the problem of joint mitigation of structured interference and non-Gaussian ambient noise is of interest [6]. A recent study [7] has shown that the performance gains

---

[*] This work was supported by the U.S. National Science Foundation under Grant NCR-9613108.

afforded by maximum likelihood (ML) multiuser detection in impulsive noise can be substantial when compared to optimum multiuser detection based on a Gaussian noise assumption. However, the computational complexity of ML detection is quite high, and therefore effective near-optimal multiuser detection techniques in non-Gaussian noise are needed. In this paper, we consider problem of mitigating multiple-access interference (MAI) in direct-sequence (DS) code-division multiple-access (DS-CDMA) channels with non-Gaussian ambient noise.

2 System Model

Consider a baseband digital DS-CDMA network operating with a coherent BPSK modulation format. The waveform received by a given terminal in such a network can be modeled as consisting of a set of superimposed modulated data signals observed in additive noise:

$$r(t) = S(t) + n(t), \qquad -\infty < t < \infty, \tag{1}$$

where $S(t)$ and $n(t)$ represent the useful signal and the ambient channel noise, respectively. The useful signal $S(t)$ is comprised of the data signals of K active users in the channel, and can be written as

$$S(t) = \sum_{k=1}^{K} A_k \sum_{i=0}^{M-1} b_k(i) s_k(t - iT - \tau_k), \tag{2}$$

where M is the number of data symbols per user in the data frame of interest, T is the symbol interval, and where A_k, τ_k, $\{b_k(i);\ i = 0, 1, \cdots, M-1\}$, and $\{s_k(t);\ 0 \le t \le T\}$, denote, respectively, the received amplitude, delay, symbol stream, and normalized signaling waveform of the k-th user. It is assumed that $s_k(t)$ is supported only on the interval $[0, T]$ and has unit energy, and that $\{b_k(i)\}$ is a collection of independent equiprobable ± 1 random variables. For the direct-sequence spread-spectrum (DS-SS) multiple access format, the user signaling waveforms are of the form

$$s_k(t) = \sum_{j=0}^{N-1} \beta_j^k \varphi(t - jT_c), \qquad t \in [0, T], \tag{3}$$

where N is the processing gain; $\left(\beta_0^k, \beta_1^k, \cdots, \beta_{N-1}^k\right)$ is a signature sequence of ± 1's assigned to the k-th user; and φ is a normalized chip waveform of duration T_c, where $NT_c = T$.

For the sake of simplicity of discussion, when developing the robust multiuser detection algorithms in the next two sections, we restrict our attention to the synchronous case of model (2), in which $\tau_1 = \tau_2 = \cdots = \tau_K = 0$. This does not incur any loss of generality, as the techniques developed for synchronous systems can be readily applied to the asynchronous systems with properly windowed received signal. For the synchronous case of model (2), to demodulate the i-th symbols of the K users, $\{b_k(i)\}_{k=1}^{K}$, it is sufficient to consider the received signal during the i-th signaling interval, i.e.,

$$r(t) = \sum_{k=1}^{K} A_k \, b_k(i) \, s_k(t - iT) + n(t), \qquad t \in [iT, (i+1)T). \tag{4}$$

At the receiver, the received signal $r(t)$ is first filtered by a chip-matched filter and then sampled at the chip rate. The resulting discrete-time signal sample at the j-th chip period of the i-th symbol interval is given by

$$r_j(i) \triangleq \int_{iT+jT_c}^{iT+(j+1)T_c} r(t)\,\varphi(t-iT-jT_c)\,\mathrm{d}t \;=\; \sum_{k=1}^{K} A_k\, b_k(i)\, s_j^k + n_j(i), \tag{5}$$

or in vector form

$$\underline{r}(i) = \sum_{k=1}^{K} A_k b_k(i)\, \underline{s}_k + \underline{n}(i), \tag{6}$$

where $\underline{s}_k \triangleq [s_0^k \;\cdots\; s_{N-1}^k]^T = \frac{1}{\sqrt{N}}[\beta_0^k \;\cdots\; \beta_{N-1}^k]^T$, is the normalized signature sequence of the k-th user, and $\underline{n}(i) \triangleq [n_0(i) \;\cdots\; n_{N-1}(i)]^T$ is the channel ambient noise sample vector at the i-th symbol interval. It is assumed that the sequence of noise samples $\{n_j(i)\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with a non-Gaussian distribution.

In this paper, we adopt the commonly used two-term Gaussian mixture model for the additive noise samples $\{n_j(i)\}$. The probability density function (pdf) of this noise model has the form

$$f = (1-\epsilon)\,\mathcal{N}\left(0, \nu^2\right) + \epsilon\,\mathcal{N}\left(0, \kappa\nu^2\right), \tag{7}$$

with $\nu > 0$, $0 \le \epsilon \le 1$, and $\kappa \ge 1$. Here the $\mathcal{N}(0, \nu^2)$ term represents the nominal background noise, and the $\mathcal{N}(0, \kappa\nu^2)$ term represents the impulsive component, with ϵ representing the probability that impulses occur. It is usually of interest to study the effects of variation in the shape of a distribution on the performance of the system, by varying the parameters ϵ and κ with fixed total noise variance

$$\sigma^2 \triangleq (1-\epsilon)\,\nu^2 + \epsilon\,\kappa\,\nu^2. \tag{8}$$

3 Multiuser Detection via Robust Regression

3.1 Least Squares Regression and the Linear Decorrelator

Consider the synchronous signal model (6). For simplicity we drop the symbol index i and denote $\theta_k \triangleq A_k b_k$. Then (6) can be rewritten as

$$r_j \;=\; \sum_{k=1}^{K} s_j^k\, \theta_k + n_j, \qquad j = 1, \cdots, N, \tag{9}$$

or in matrix notation,

$$\underline{r} \;=\; \underline{S}\,\underline{\theta} + \underline{n}, \tag{10}$$

where $\underline{S} \triangleq [\underline{s}_1\, \underline{s}_2 \cdots \underline{s}_K]$, and $\underline{\theta} \triangleq [\theta_1\, \theta_2 \cdots \theta_K]^T$. Consider the linear regression problem of estimating the K unknown parameters $\theta_1, \theta_2, \cdots, \theta_K$ from the N observations

$r_1, r_2, \cdots, r_N$ in (9). Classically this problem can be solved by minimizing the sum of squared errors, i.e., through the least-squares (LS) method:

$$\hat{\underline{\theta}}_{LS} = \arg\min_{\underline{\theta}} \sum_{j=1}^{N} \left(r_j - \sum_{k=1}^{K} s_j^k\, \theta_k \right)^2 = \arg\min_{\underline{\theta}} \left\| \underline{r} - \underline{S}\, \underline{\theta} \right\|^2 . \tag{11}$$

If $n_j \sim \mathcal{N}(0,\ \sigma^2)$, then the probability density function (pdf) of the received signal $\underline{r}$ under the true parameters $\underline{\theta}$ is given by

$$f_{\underline{\theta}}(\underline{r}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{\left\| \underline{r} - \underline{S}\, \underline{\theta} \right\|^2}{2\sigma^2} \right). \tag{12}$$

It is easily seen from (12) that the maximum likelihood estimate of $\underline{\theta}$ under the i.i.d. Gaussian noise samples is given by the LS solution $\hat{\underline{\theta}}_{LS}$ in (11). Upon differentiating (11), $\hat{\underline{\theta}}_{LS}$ is then the solution to the following linear system equations

$$\sum_{j=1}^{N} \left(r_j - \sum_{k=1}^{K} s_j^k\, \theta_k \right) s_j^k = 0, \qquad j = 1, \cdots, N, \tag{13}$$

or in matrix form

$$\underline{S}^T \underline{S}\, \underline{\theta} = \underline{S}^T \underline{r}. \tag{14}$$

Define the cross-correlation matrix of the signature waveforms of all users as $\underline{R} \triangleq \underline{S}^T \underline{S}$. Assuming that the user signature waveforms are linearly independent, i.e., $\underline{S}$ has a full rank K, then $\underline{R}$ is invertible, and the LS solution to (13) or (14) is given by

$$\hat{\underline{\theta}}_{LS} = \left(\underline{S}^T \underline{S} \right)^{-1} \underline{S}^T \underline{r} = \underline{R}^{-1} \underline{S}^T \underline{r}. \tag{15}$$

We observe from (15) that the LS estimate $\hat{\underline{\theta}}_{LS}$ is exactly the output of the linear decorrelating multiuser detector for the K users [3]. This is not surprising, since the linear decorrelating detector gives the maximum likelihood estimate of the product of the amplitude and the data bit $\theta_k = A_k b_k$ in Gaussian noise [3]. Given the estimate $\hat{\theta}_k$, the estimated amplitude and the data bit are then determined by

$$\hat{A}_k = |\hat{\theta}_k|, \qquad \hat{b}_k = \operatorname{sgn}(\hat{\theta}_k). \tag{16}$$

3.2 Robust Multiuser Detection via M-Estimation-based Regression

It is well known that the LS estimate is very sensitive to the tail behavior of the noise density. Its performance depends significantly on the Gaussian assumption and even a slight deviation of the noise density from the Gaussian distribution can, in principle, cause a substantial degradation of the LS estimate. Since the linear decorrelating multiuser detector is in the form of the LS solution to a linear regression problem, it can be concluded that its performance is also sensitive to the tail behavior of the noise distribution. As will be demonstrated in Section 3.5, the performance of the linear decorrelating detector degrades substantially if the ambient noise deviates even slightly from a Gaussian distribution. In this section, we consider some robustized versions of the decorrelating multiuser detector, which are nonlinear in nature. Robustness of an estimator refers to its insensitivity to small changes in the underlying statistical model.

The LS estimate corresponding to equations (11) and (13) can be robustized by using the class of M-estimators proposed by Huber [2]. Instead of minimizing a sum of squared residuals as in (11), Huber proposed to minimize a sum of a less rapidly increasing function, ρ, of the residuals:

$$\hat{\underline{\theta}} \;=\; \arg\min_{\underline{\theta}\in\mathbb{R}^K} \sum_{j=1}^{N} \rho\left(r_j - \sum_{k=1}^{K} s_j^k\,\theta_k\right). \tag{17}$$

Suppose that ρ has a derivative $\psi = \rho'$, then the solution to (17) satisfies the implicit equation

$$\sum_{j=1}^{N} \psi\left(r_j - \sum_{k=1}^{K} s_j^k\,\theta_k\right) s_j^k \;=\; 0, \qquad j = 1, \cdots, N, \tag{18}$$

or in vector form

$$\underline{S}^T\, \psi\,(\underline{r} - \underline{S}\,\underline{\theta}) = \underline{0}_K, \tag{19}$$

where $\psi(\underline{x}) \triangleq [\psi(x_1), \cdots, \psi(x_K)]^T$ for any $\underline{x} \in \mathbb{R}^K$; and $\underline{0}_K$ denotes a K-dimensional zero vector. An estimator defined by (17) or (18) is called an *M-estimator.* The name "M-estimator" comes from "maximum likelihood type estimator" [2], since the choice of $\rho(x) = -\log f(x)$ gives the ordinary maximum likelihood estimates. If ρ is convex, then (17) and (18) are equivalent; otherwise (17) is still very useful in searching for the solution to (18). To achieve robustness, it is necessary that ψ be bounded and continuous. Usually to achieve high efficiency at the Gaussian case, we require that $\psi(x) \approx x$ for x small. Consistency of the estimate requires that $E\,\psi(n_j) = 0$. Hence for symmetric noise densities, ψ is usually odd symmetric. We next consider some specific choices of the penalty function ρ and the corresponding derivative ψ.

1. Linear decorrelating detector

The linear decorrelating detector, which is simply the LS estimator, corresponds to having the penalty function and its derivative, respectively, as

$$\rho_{LS}(x) = \frac{x^2}{2\gamma}, \qquad \psi_{LS}(x) = \frac{x}{\gamma}, \tag{20}$$

where γ is any positive constant. Notice that the linear decorrelating detector is scale invariant.

2. Maximum likelihood decorrelating detector

The maximum likelihood decorrelating detector in the non-Gaussian noise with pdf f [in the sense that it is a maximum likelihood estimate of the product of the amplitude and data bit $\theta_k \triangleq A_k b_k$] is given by the M-estimator with the penalty function and its derivative, respectively, as

$$\rho_{ML}(x) = -\log f(x), \qquad \psi_{ML}(x) = -\frac{f'(x)}{f(x)}. \tag{21}$$

3. Minimax decorrelating detector

Assuming that the sample distribution belongs to the set of *ϵ-contaminated Gaussian models*, given by

$$\mathcal{P}_\epsilon \;\triangleq\; \{(1-\epsilon)\mathcal{N}(0,\nu^2)+\epsilon H;\, H \text{ is a symmetric distribution}\}, \tag{22}$$

where $0<\epsilon<1$ is fixed, and ν^2 is the variance of the nominal Gaussian distribution. The minmax robust decorrelating detector essentially gives the maximum likelihood estimate of $\underline{\theta}$ under the least favorable pdf in $\mathcal{P}_\epsilon$. The corresponding minimax M-estimator is then determined by the Huber penalty function and its derivative [2], given respectively by

$$\rho_H(x)=\begin{cases}\frac{x^2}{2\nu^2}, & \text{for } |x|\le k\nu^2,\\ \frac{k^2\nu^2}{2}-k|x|, & \text{for } |x|>k\nu^2,\end{cases}\qquad \psi_H(x)=\begin{cases}\frac{x}{\nu^2}, & \text{for } |x|\le k\nu^2,\\ k\,\mathrm{sgn}(x), & \text{for } |x|>k\nu^2,\end{cases} \tag{23}$$

where k, ϵ and ν are connected through

$$\frac{\phi(k\nu)}{k\nu}-Q(k\nu) \;=\; \frac{\epsilon}{2(1-\epsilon)}, \tag{24}$$

where $\phi(x)\triangleq\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, and $Q(x)\triangleq\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{x^2}{2}}dx$. The minimax robust decorrelating detector is obtained by substituting ρ_H and ψ_H into (17) and (18). The nonlinearities ψ for the three types of decorrelating detectors are plotted in Figure 1 for $\epsilon=0.01$.

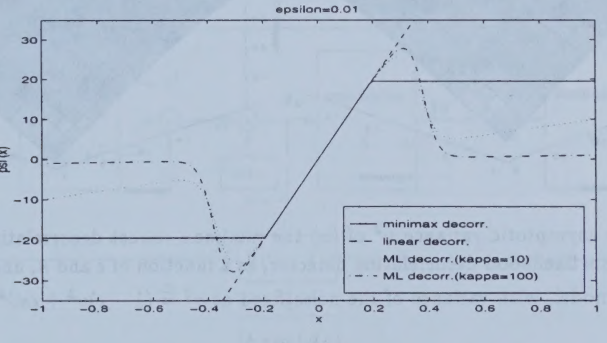


Figure 1: The nonlinearity ψ. ($\nu^2=0.01$).

3.3 Asymptotic Performance of Robust Multiuser Detector

Assuming that the signature sequences of all users are random, we consider the asymptotic probability of error for the class of decorrelating detectors defined by (18), for large processing gain $N\longrightarrow\infty$. It can be shown that the asymptotic probability of error for the k-th user is given by

$$P_e^k \;\triangleq\; Pr\left(\hat{\theta}_k<0\mid\theta_k>0\right) \;=\; Q\left(\frac{A_k}{\upsilon\sqrt{[\underline{R}]^{-1}_{kk}}}\right), \tag{25}$$

where $\underline{R}$ is the cross-correlation matrix of the signature waveforms of all users, and υ is the asymptotic variance given by

$$\upsilon^2 \triangleq \frac{\int \psi^2(u) f(u) \mathrm{d}u}{\left[\int \psi'(u) f(u) \mathrm{d}u\right]^2}. \tag{26}$$

Hence for the class of M-decorrelators defined by (18), their asymptotic probabilities of detection error are connected through the parameter υ. In Figure 2 we plot the asymptotic variance υ^2 of the maximum likelihood decorrelator and the minimax robust decorrelator as a function of ϵ and κ, under the Gaussian mixture noise model (7). The total noise variance is kept constant as ϵ and κ vary, i.e., $\sigma^2 \triangleq (1-\epsilon)\nu^2 + \epsilon\kappa\nu^2 = (0.1)^2$. From the two plots we see that the two nonlinear detectors have very similar asymptotic performance. Moreover, in this case the asymptotic variance υ^2 is a decreasing function of either ϵ or κ when one of them is fixed. The asymptotic variance of both the nonlinear decorrelators are strictly less than that of the linear decorrelator, which corresponds to a plane at $\upsilon^2 = \sigma^2 = (0.1)^2$.

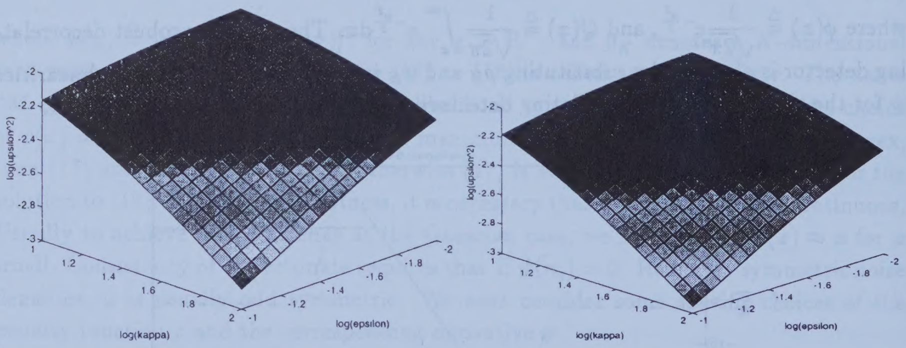


Figure 2: The asymptotic variance υ^2 of (a) the minimax robust decorrelating detector, and (b) the maximum likelihood decorrelating detector, as a function of ϵ and κ, under the Gaussian mixture noise model, with variance of the noise fixed at $\sigma^2 \triangleq (1-\epsilon)\nu^2 + \epsilon\kappa\nu^2 = (0.1)^2$.

3.4 Computational Issues

Assume that the penalty function $\rho(x)$ in (17) has a bounded second order derivative, i.e., $|\rho''(x)| = |\psi'(x)| \leq \mu$, for some $\mu > 0$. Then equation (18) can be solved iteratively by the following modified residual method [2]. Let $\underline{\theta}^l$ be the estimate at the l-th step, then it is updated according to

$$\underline{z}^l \triangleq \psi\left(\underline{r} - \underline{S}^T \underline{\theta}^l\right), \tag{27}$$

$$\underline{\theta}^{l+1} = \underline{\theta}^l + \frac{1}{\mu}\left(\underline{S}^T \underline{S}\right)^{-1} \underline{S}^T \underline{z}^l, \tag{28}$$

where μ is a step-size parameter. Denote the cost function in (17) by $C(\underline{\theta}) \triangleq \sum_{j=1}^{N} \rho\left(r_j - \underline{\xi}_j^T \underline{\theta}\right)$. We have the following result regarding the convergence behavior of the above iterative procedure.

Proposition 1 *If $|\psi'(x)| \le \mu$, then the iterative estimates defined by (27) and (28) satisfy*

$$C\left(\underline{\theta}^l\right) - C\left(\underline{\theta}^{l+1}\right) \ge \frac{\mu}{2}\left(\underline{\theta}^l - \underline{\theta}^{l+1}\right)^T \underline{R}\left(\underline{\theta}^l - \underline{\theta}^{l+1}\right) = \frac{1}{2\mu}\underline{z}\left(\underline{\theta}^l\right)^T \underline{S}\,\underline{R}^{-1}\underline{S}^T \underline{z}\left(\underline{\theta}^l\right), \quad (29)$$

where $\underline{R} \triangleq \underline{S}^T\underline{S}$, is assumed to be positive definite; and $\underline{z}(\underline{\theta}) \triangleq \psi(\underline{r} - \underline{S}^T\underline{\theta})$. Furthermore, if $\rho(x)$ is convex and bounded from below, then $\underline{\theta}^l \longrightarrow \underline{\theta}^*$ as $l \longrightarrow \infty$, where $\underline{\theta}^*$ is the unique minimum point of the cost function $C(\underline{\theta})$ (i.e., the unique solution to (18)).

Notice that matrix $\frac{1}{\mu}(\underline{S}^T\underline{S})^{-1}\underline{S}^T$ in (28) can be computed offline, and the major computation involved at each iteration is the product of this $K \times K$ matrix with a K-vector $\underline{z}^l$. For the initial value $\underline{\theta}^0$ we can take the least squares solution, i.e.,

$$\underline{\theta}^0 = \frac{1}{\mu}\left(\underline{S}^T\underline{S}\right)^{-1}\underline{S}^T\underline{r}. \quad (30)$$

The iteration is stopped if $\|\underline{\theta}^l - \underline{\theta}^{l-1}\| \le \varepsilon$, for some small number ε. The numerical experiments show that on average it takes less than 10 iterations for the algorithm to converge.

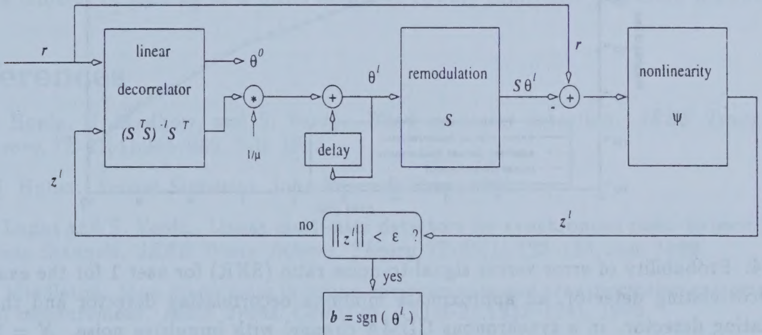


Figure 3: Diagram of the M-decorrelating multiuser detector, which is a robustized version of the linear decorrelating multiuser detector.

The operations of the M-decorrelating multiuser detector are depicted in Figure 3. It is evident that it is essentially a robustized version of the linear decorrelating detector. At each iteration, the residual signal, which is the difference between the received signal $\underline{r}$ and the remodulated signal $\underline{S}\,\underline{\theta}^l$, is passed through the nonlinearity $\psi(\cdot)$. Then the modified residual $\underline{z}^l$ is passed through the linear decorrelating filter to get the modification on the previous estimate.

3.5 Simulation Examples

We demonstrate the performance gains achieved by the minimax robust decorrelating detector over the linear decorrelator in impulsive noise. The noise distribution parameters are $\epsilon = 0.01$ and $\kappa = 100$. The bit error rate versus SNR for the two detectors is plotted in Figure 4. Also shown in this figure is the performance of an "approximate" minimax decorrelating detector, in which the parameter k is taken as $k = \frac{1.5}{\sigma}$, and the step size parameter μ is set as $\mu = \sigma^2$. The reason for studying such an approximate robust detector is that in practice, it is unlikely that the exact parameters ϵ and ν in the noise model (7) are known to the receiver. However, the total noise variance σ^2 can be estimated from the received signal (as discussed in the next section). Hence if we could set some simple rule for choosing the parameters k [as opposed to calculating it exactly from equation (24)] and μ, then this approximate robust detector is much easier to implement than the exact one. It is seen from Figure 4 that the robust decorrelating multiuser detector offers significant performance gains over the linear decorrelating detector. Moreover this performance gain increases as the SNR increases. Another important observation is that the performance of the robust multiuser detector is insensitive to the parameters ϵ and κ in the noise model, which is evidenced by the fact that the performance of the approximate robust detector is very close to that of the exact robust detector.

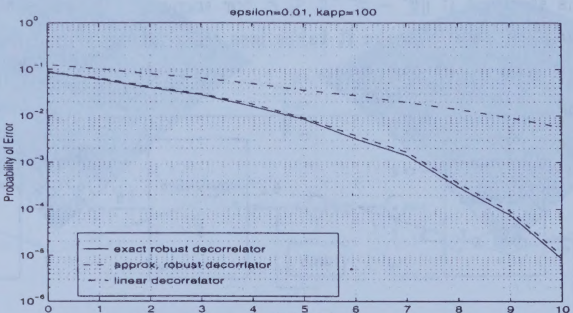


Figure 4: Probability of error versus signal-to-noise ratio (SNR) for user 1 for the exact minimax decorrelating detector, an approximate minimax decorrelating detector and the linear decorrelating detector. in a synchronous CDMA channel with impulsive noise. $N = 31, K = 6, \epsilon = 0.01, \kappa = 100$. The powers of the interferers are 10dB above the power of user 1.

4 Discussions and Conclusion

In many practical wireless channels in which multiuser detection techniques may be applied, the ambient noise is likely to have an impulsive component that gives rise to larger tail probabilities than is predicted by the Gaussian model. Impulsive noise can seriously degrade the error probability of the linear multiuser detectors for a given level of the noise power. In this paper, we have proposed a robust multiuser detection technique that is seen to significantly outperform the linear multiuser detectors in non-Gaussian ambient noise. This technique is based on the M-estimation method for robust regression. The asymptotic performance of several versions of the robust detectors are analyzed and compared. Simulations show that the robust multiuser detectors offer significant performance gain over the linear multiuser detectors in non-Gaussian noise, and they incur little performance loss when the channel noise is indeed Gaussian.

Throughout this paper, we have assumed that the signature waveforms of all users are known to the receiver in order to implement the robust multiuser detectors. One remarkable feature of the linear multiuser detectors is that there exist *blind* techniques that can be used to adapt these detectors, which allow one to use a linear multiuser detector for a given user with no knowledge beyond that required for implementation of the conventional matched-filter detector for that user.

There are two major approaches to blind adaptive multiuser detection. In the first approach the received signal is passed through a linear filter, which is chosen to minimize, within a constraint, the mean-square value of its output [1]. Adaptation algorithms such as least-mean-squares (LMS) or recursive-least-squares (RLS) can be applied for updating the filter weights. Ideally the adaptation will lead the filter converge to the linear MMSE multiuser detector, irrespective of the noise distribution. [In practice the impulsiveness of the noise will slow down the convergence]. Therefore this approach can not be used to adapt the robust multiuser detector, which is nonlinear in nature.

Another approach to blind multiuser detection is the subspace-based method proposed in [8, 9], through which both the linear decorrelating detector and the linear MMSE detector can be obtained blindly. As will be discussed in a forthcoming paper, this approach is more fruitful in leading to a blind adaptive robust multiuser detection method.

References

[1] M. Honig, U. Madhow, and S. Verdú. Blind multiuser detection. *IEEE Trans. Inform. Theory*, IT-41(4):944–960, July 1995.

[2] P.J. Huber. *Robust Statistics.* John Wiley & Sons, 1981.

[3] R. Lupas and S. Verdú. Linear multi-user detectors for synchronous code-division multiple-access channels. *IEEE Trans. Inform. Theory*, IT-35(1):123–136, Jan. 1989.

[4] D. Middleton. Man-made noise in urban environments and transportation systems: models and measurement. *IEEE Trans. Commun.*, COM-21:1232–1241, 1973.

[5] D. Middleton. Statistical-physical models of electromagnetic interference. *IEEE Trans. Electromag. Compat.*, EMC-19:106–127, 1977.

[6] H.V. Poor. Non-Gaussian signal processing problems in multiple-access communications. In *Proc. 1996 USC/CRASP Workshop on Non-Gaussian Signal Processing*, Ft. George Meade, MD, May 1996.

[7] H.V. Poor and M. Tanda. An analysis of some multiuser detectors in impulsive noise. In *Proc. 16th GRETSI Symposium on Signal and Image Processing*, Grenoble, France, Sept. 1997.

[8] X. Wang and H.V. Poor. Blind multiuser detection: A subspace approach. To appear in *IEEE Trans. Inform. Theory.*

[9] X. Wang and H.V. Poor. Blind adaptive interference suppression for CDMA communications based on eigenspace tracking. In *Proc. 1997 Conference on Information Sciences and Systems*, Baltimore, MD, Mar. 1997.

An Optimal One-shot Multiuser Detector for Flat Rayleigh Fading Channels

Andrew Sendonaris and Behnaam Aazhang

sendos@rice.edu aaz@rice.edu

Laboratory for Communications Research,
Department of Electrical and Computer Engineering,
Rice University, Houston, TX.

Abstract

In this study[1] we derive the general form of an optimal one-shot multiuser detector given any density for the users' amplitudes. It is shown that well-known multiuser receivers such as the maximum likelihood detector, the decorrelating detector and even the MMSE detector are special cases of the above receiver, each optimal given a particular assumption about the density of the users' amplitudes. Finally, when it is assumed that the amplitudes obey a Rayleigh distribution, we obtain the optimal one-shot multiuser detector for flat Rayleigh fading channels. Results show that, as expected, the resulting receiver has superior performance.

I INTRODUCTION

With the advent of multiuser detectors [1], CDMA has proven to be a viable technology for the next generation of wireless systems (see, for example, [2]). However, one of the many challenges unique to the wireless channel, and one that has not been fully resolved, is the fact that the multipath fading environment is time-varying. Even though the statistics of the channel variations may be known, most multiuser detectors do not incorporate this knowledge into the detection process. Any such knowledge is most often used for tracking the channel variations. The minimum P_e (probability of bit error) receiver, however, is a MAP (Maximum a-posteriori) receiver, that is, a receiver that directly incorporates the density of the channel variations into the bit decisions.

To explore the potential gains of a multiuser MAP receiver, this study focuses on flat fading environments. The attenuation suffered by the transmitted signal is well-modeled as being composed of three components: deterministic path loss, large (time) scale fading, and small (time) scale fading. The latter component is well-modeled as a Rayleigh random variable. Therefore, over a small time scale, since the deterministic and the large time scale components remain approximately constant, the received amplitude may be modeled as a Rayleigh r.v. with a fixed mean. This mean is time-varying, but only over larger time scales, thus making it easier to track by the receiver.

---

[1] Supported in part by a grant from the Texas Advanced Technology Program and the NSF

Thus, based on the assumption that the receiver tracks the slowly varying means of the users' received amplitudes, this work addresses the issue of finding the optimal multiuser one-shot detector, that is, a detector that forms its bit decisions using only the received signal in the corresponding symbol period. Possible reasons for using a one-shot detector may be a channel is that is fading too rapidly, or a desire to reduce the complexity of the receiver. Given the fact that the detector is one-shot, it is necessary to assume that the there is a phase-tracking algorithm in operation [3], due to the phase ambiguity that would otherwise arise.

It should be noted that other work has also dealt with the issue of MAP (i.e. minimum P_e) receivers in Rayleigh fading environments [4–9], but the work focused on single user systems and/or on (computationaly complex) sequence detection.

II PROBLEM SETUP

We assume a synchronous CDMA system that suffers from single-path Rayleigh fading and AWGN. Let N_c denote the spreading factor (number of chips per bit) and assume there are K users in the system. Then, the resulting received $N_c \times 1$ vector, after chip-matched filtering, is given by

$$\mathbf{r} = \mathbf{C}\,\mathbf{W}\,\mathbf{b} + \mathbf{n} \tag{1}$$

where $\mathbf{C}$ is an $N_c \times K$ matrix whose columns correspond to the spreading codes of the users, $\mathbf{W}$ is a $K \times K$ diagonal matrix whose entries correspond to the (complex) received amplitudes of the various users, $\mathbf{b}$ is a $K \times 1$ vector containing the users' bits, and $\mathbf{n}$ is an $N_c \times 1$ zero-mean complex Gaussian vector with covariance matrix equal to $\sigma^2 I_K$, where I_K is the $K \times K$ identity matrix.

In order to facilitate analysis, $\mathbf{W}$ may be re-written as $\mathbf{W} = \Phi\,\mathbf{A}$ where Φ and $\mathbf{A}$ are diagonal matrices that contain the phases and the (positive real) amplitudes, respectively, of the various users. Based on the discussion in Section I, Φ is assumed to be known and can thus be combined with $\mathbf{C}$ to form $\bar{\mathbf{C}} = \mathbf{C}\,\Phi$. As a result, (1) may be written as

$$\mathbf{r} = \bar{\mathbf{C}}\,\mathbf{A}\,\mathbf{b} + \mathbf{n} \tag{2}$$

Based on the above system model, the optimal estimator-detector is given by

$$\begin{aligned} \hat{\mathbf{A}}, \hat{\mathbf{b}} &= \arg \max_{\substack{\mathbf{A}\in\mathbb{R}_+^{K\times K} \\ \mathbf{b}\in\{-1,1\}^K}} p_{A,b|R}(\mathbf{A}, \mathbf{b}|\mathbf{r}) \\ &\equiv \arg \max_{\substack{\mathbf{A}\in\mathbb{R}_+^{K\times K} \\ \mathbf{b}\in\{-1,1\}^K}} p_{R|A,b}(\mathbf{r}|\mathbf{A}, \mathbf{b})\, p_{A,b}(\mathbf{A}, \mathbf{b}) \end{aligned} \tag{3}$$

where $\mathbb{R}_{+}^{K\times K}$ denotes the set of all $K \times K$ diagonal matrices with positive real entries. The density of $\mathbf{r}$, conditioned on the amplitudes and bits, is

$$\begin{aligned} p_{R|A,b}(\mathbf{r}|\mathbf{A},\mathbf{b}) &= p_N(\mathbf{r} - \bar{\mathbf{C}}\,\mathbf{A}\,\mathbf{b}) \\ &= \frac{1}{(2\pi\sigma^2)^{N_c/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{r} - \bar{\mathbf{C}}\,\mathbf{A}\,\mathbf{b})^H(\mathbf{r} - \bar{\mathbf{C}}\,\mathbf{A}\,\mathbf{b})\right) \end{aligned} \tag{4}$$

Also, assuming that the users' bits are equiprobable, that the users' received amplitudes are statistically independent, and that the users' received amplitudes are statistically independent with the users' bits, we have

$$p_{A,b}(\mathbf{A},\mathbf{b}) = \left(\frac{1}{2}\right)^K \prod_{i=1}^{K} p_{A_{i,i}}(\mathbf{A}_{i,i}) \tag{5}$$

where $p_{A_{i,i}}$ denotes the density of the received amplitude of the i^{th} user. Consequently, the optimization problem in (3) becomes

$$\hat{\mathbf{A}},\hat{\mathbf{b}} = \arg \min_{\substack{\mathbf{A}\in\mathbb{R}_{+}^{K\times K} \\ \mathbf{b}\in\{-1,1\}^K}} \frac{1}{2}\mathbf{b}^\top\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{b} - \mathbf{b}^\top\mathbf{A}^\top\mathbf{y}_{MF} - \sigma^2\sum_{i=1}^{K}\ln p_{A_{i,i}}(\mathbf{A}_{i,i}) \tag{6}$$

where $\mathbf{R} = \mathrm{Re}\left[\bar{\mathbf{C}}^H\bar{\mathbf{C}}\right]$ is the correlation matrix of the users' codes and $\mathbf{y}_{MF} = \mathrm{Re}\left[\bar{\mathbf{C}}^H\mathbf{r}\right]$ is the vector of matched-filter outputs. It appears that the problem in (6) is not only analytically intractable, but also prohibitively computationally intensive. This difficulty is alleviated if the problem is re-cast as follows.

Since the elements of $\mathbf{A}$ are positive, $\mathbf{A}$ may be combined with $\mathbf{b}$ to form $\mathbf{x} = \mathbf{A}\,\mathbf{b} \in \mathbb{R}^K$. No information is lost in doing so since the bits and amplitudes may be extracted from $\mathbf{x}$ according to $\mathbf{b}_i = \mathrm{sign}(\mathbf{x}_i)$ and $\mathbf{A}_{i,i} = |\mathbf{x}_i|$. Given the above definition of $\mathbf{x}$, it can be shown that $p_{\mathbf{x}_i}(x_i) = \frac{1}{2}p_{A_{i,i}}(|x_i|)$ and that (6) becomes

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^K} \frac{1}{2}\mathbf{x}^\top\mathbf{R}\mathbf{x} - -\mathbf{x}^\top\mathbf{y}_{MF} - \sigma^2\sum_{i=1}^{K}\ln p_{\mathbf{x}_i}(x_i) \tag{7}$$

with $\hat{\mathbf{b}}_i = \mathrm{sign}(\hat{\mathbf{x}}_i)$ and $\hat{\mathbf{A}}_{i,i} = |\hat{\mathbf{x}}_i|$. The first-order conditions for the above problem are given by

$$\mathbf{R}\mathbf{x} - \mathbf{y}_{MF} - \sigma^2\left[\frac{p'_{\mathbf{x}_i}(x_i)}{p_{\mathbf{x}_i}(x_i)}\right] = 0 \tag{8}$$

Equations (7) and (8) define the optimal one-shot receiver in a flat fading environment.

III COMPARISON OF AMPLITUDE DENSITY ASSUMPTIONS

The analysis in section II gives us a class of optimal receivers: each member of this class corresponds to a particular assumption about the statistics of the amplitude variations of the various users. Section IV studies the optimal one-shot multiuser detector in a flat Rayleigh fading channel, that is, when it is assumed that $p_{A_{i,i}}$ is a Rayleigh density. However, for greater insight, this Section investigates three different assumptions about $p_{A_{i,i}}$, all of which lead to well-known multiuser receivers. In fact, due to the fact that these receivers are well-known, we will not elaborate on their properties.

III.1 Approach 1: Assume perfect knowledge of amplitudes

Under this approach, $p_{X_i}(x_i)$ consists of two delta functions, one at $-A_i$ and one at $+A_i$. Therefore, the problem in (7) reduces to

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}\in\{-A_i,+A_i\}^K} \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{R}\mathbf{x} - \mathbf{x}^{\mathsf{T}}\mathbf{y}_{MF} \tag{9}$$

The above problem has no closed-form solution and thus it is necessary to search over all combinations of $+/-A_i$'s. This results in what is known as the "Maximum Likelihood Detector".

III.2 Approach 2: Assume no knowledge of amplitudes

This assumption corresponds to a "uniform" $p_{X_i}(x_i)$, that is, we do not assume that any value of x_i is more likely than any other. As a result, (7) reduces to

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}\in\mathbb{R}^K} \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{R}\mathbf{x} - \mathbf{x}^{\mathsf{T}}\mathbf{y}_{MF} \tag{10}$$

whose solution is easily shown to be

$$\mathbf{x} = \mathbf{R}^{-1}\mathbf{y}_{MF} \tag{11}$$

This is of course what is known as the "Decorrelating detector".

III.3 Approach 3: Assume amplitude density is a one-sided Gaussian

The reason for this peculiar choice of amplitude density will become apparent shortly. Under this assumption, $p_{X_i}(x_i)$ is a zero-mean Gaussian density with variance σ_i^2. Consequently, the first-order conditions in (8) become

$$(\mathbf{R} + \mathrm{diag}(\frac{\sigma^2}{\sigma_i^2}))\mathbf{x} - \mathbf{y}_{MF} = 0 \tag{12}$$

Thus, if we set $\sigma_i = A_{i,i}$ we obtain

$$\mathbf{x} = (\mathbf{R} + \mathrm{diag}(\frac{\sigma^2}{A_{i,i}^2}))-1\mathbf{y}_{MF} \tag{13}$$

which is of course the "MMSE detector". We therefore see that, even though the MMSE is known to be optimal in the minimum mean square error sense, it is also optimal in the minimum P_e sense, given a particular assumption about the density of the users' amplitudes.

III.4 Optimality considerations

Even though it is interesting and insightful to see how the previous three well-known receivers may be looked at as optimal receivers under different amplitude density assumptions, the fact is that there are problems with all three of them. First, it is too optimistic to assume perfect amplitude knowledge, and as a result, receiver performance will suffer in the presence of estimation errors. Second, it is too pessimistic to assume no amplitude knowledge, since this degrades receiver performance. Third, the assumption of a one-sided Gaussian density for the amplitudes is ad-hoc, and does not correspond to any practical scenario.

Which density *does* correspond to many practical scenarios is known. Section IV analyzes the receiver that results from the incorporation of this knowledge into (7).

IV OPTIMAL RECEIVER

As mentioned in Section I, a user's received amplitude is well modeled as a Rayleigh r.v. with a mean that remains approximately constant over small time scales. Therefore, the model we use for $p_{x_i}(x_i)$ is

$$p_{x_i}(x_i) = \frac{\pi}{4m_i^2}|x_i|e^{-\frac{\pi}{4m_i^2}x_i^2} \tag{14}$$

where m_i is the mean of the amplitude of user i over the time period of interest. Consequently, the first-order conditions in (8) become

$$(\mathbf{R} + \mathrm{diag}(\frac{\sigma^2}{\frac{2}{\pi}m_i^2}))\mathbf{x} - \mathbf{y}_{MF} - \sigma^2./\mathbf{x} = 0 \tag{15}$$

where $\sigma^2./\mathbf{x}$ denotes the vector whose i^{th} component equals $\frac{\sigma^2}{\mathbf{x}_i}$. The resulting K simultaneous quadratic equations have no closed-form solution. This implies that the solution has to be found numerically. However, as can be seen from Figure 1, the objective function is not convex. Thus, even if the solution to (15) is found numerically, it will most probably only correspond to a local minimum. Consequently, arriving at the global minimum is best accomplished by numerically minimizing the objective function, rather than solving the F.O.C.'s.

Given the shape of the objective function,

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathsf{T}}(\mathbf{R} + \mathrm{diag}(\frac{\sigma^2}{\frac{2}{\pi}m_i^2}))\mathbf{x} - \mathbf{x}^{\mathsf{T}}\mathbf{y}_{MF} - \sigma^2\sum_{i=1}^{K}\ln|x_i| \tag{16}$$

that is, the fact that it is non-convex and that it approaches minus infinity as any element of $\mathbf{x}$ approaches 0, any numerical optimization algorithm would fail to find the global minimum, unless the initial guess was in the correct "quadrant". In this study, the above problem has been overcome by a two-step process. First, a smoothed version of the objective function is minimized. The intention is to get inside the quadrant that contains the global minimum. Second, the original objective function is minimized using the above solution as the starting point. Smoothing the objective function is at present achieved via an ad-hoc method. However, as Figure 1 and the results in the following section demonstrate, this two-step algorithm is effective. In addition, the computational complexity of this algorithm is manageable.

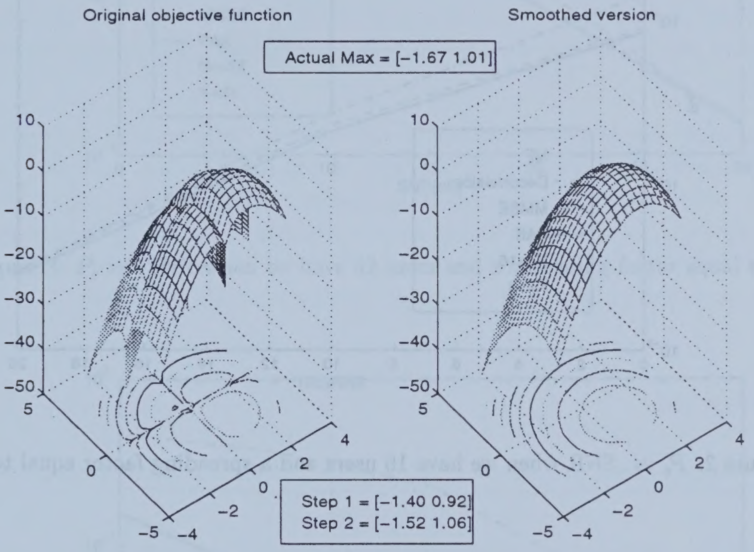


Figure 1: Demonstration of the two-step algorithm for finding the global minimum of the objective function of the optimal receiver in a flat Rayleigh fading environment. (Note: for visual clarity, what is shown is the negative of the objective function)

V RESULTS

Figures 2 through 4 show the probability of bit error of various receivers, as a function of the signal-to-noise ratio (SNR). The receivers being compared are the conventional matched filter receiver, the decorrelator, the MMSE that has perfect knowledge of users' amplitudes and the optimal receiver that has perfect knowledge of the means of the users' amplitudes. In all figures, the acronym PAK refers to the optimal receiver developed in this study, and is derived from "Partial Amplitude Knowledge" receiver, which is meant to emphasize that the receiver is incorporating partial (statistical) knowledge about the amplitudes into the decision process.

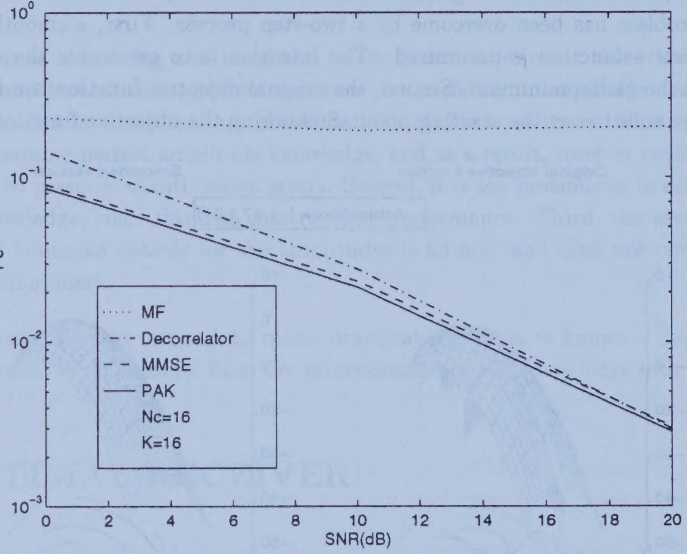


Figure 2: P_e vs. SNR when we have 16 users and a spreading factor equal to 16

VI CONCLUSIONS/DISCUSSION

We have derived the optimal one-shot multiuser detector for a flat Rayleigh fading channel. It performs significantly better than the decorrelator and is slightly better than the MMSE which assumes perfect knowledge of amplitudes.

It is important to examine what information is required in order to make the various receivers operational, besides the timing and phases of the users. The matched filter receiver requires knowledge of the desired user's spreading code. The decorrelator, MMSE and PAK receivers, as implemented in this study, require knowledge of all the users' spreading codes. However, the decorrelator requires no more than that, whereas the MMSE requires knowledge of $\frac{\sigma^2}{A_{i,i}^2}$, that is, the *instantaneous* SNR for each user. In contrast, the PAK receiver requires knowledge of m_i and σ^2, that is, the means of the users' amplitudes and the level of the background AWGN, both of which are slowly-

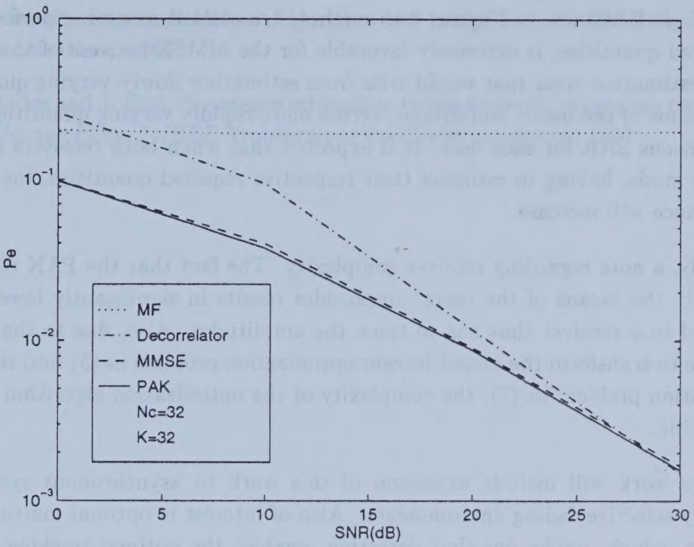


Figure 3: P_e vs. SNR when we have 32 users and a spreading factor equal to 32

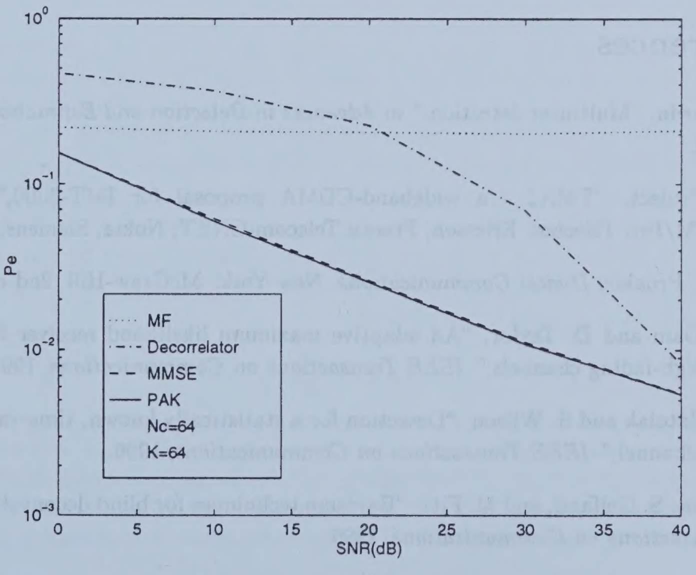


Figure 4: P_e vs. SNR when we have 64 users and a spreading factor equal to 64

varying, and can thus be easily tracked by the receiver. The fact that both the MMSE and the PAK performance, in Figures 2 through 4, are plotted assuming perfect knowledge of required quantities, is extremely favorable for the MMSE because of the significantly smaller estimation error that would arise from estimating slowly-varying quantities, such as the means of the users' amplitudes, versus more rapidly varying quantities such as the instantaneous SNR for each user. It is expected that when both receivers are run in an adaptive mode, having to estimate their respective required quantities, the difference in performance will increase.

Finally, a note regarding receiver complexity. The fact that the PAK receiver has to track only the means of the users' amplitudes results in significantly lower complexity compared to a receiver that has to track the amplitudes. Also, due to the fact that we were able to transform the mixed-integer optimization problem in (3) into the continuous optimization problem in (7), the complexity of the optimization algorithm became very manageable.

Future work will include extension of this work to asynchronous systems and to frequency-selective fading environments. Also of interest is optimal multiuser sequence detection, which, unlike one-shot detection, enables the optimal tracking of the users' phases, and which, via the exploitation of the correlation in the fading process, promises significant potential gains. More importantly, this would demonstrate if the loss in performance sustained by the use of a one-shot detector is significant or not.

References

[1] S. Verdu, "Multiuser detection," in *Advances in Detection and Estimation*, JAI Press, 1993.

[2] F. Project, "FMA2 - a wideband-CDMA proposal for IMT-2000," tech. rep., CSEM/Pro Telecom, Ericsson, France Telecom-CNET, Nokia, Siemens, 1997.

[3] J. G. Proakis, *Digital Communications.* New York: McGraw-Hill, 2nd ed., 1989.

[4] W. Dam and D. Taylor, "An adaptive maximum likelihood receiver for correlated rayleigh-fading channels," *IEEE Transactions on Communications*, 1994.

[5] D. Matolak and S. Wilson, "Detection for a statistically known, time-varying dispersive channel," *IEEE Transactions on Communications*, 1996.

[6] G. Lee, S. Gelfand, and M. Fitz, "Bayesian techniques for blind deconvolution," *IEEE Transactions on Communications*, 1996.

[7] S. Vasudevan and M. Varanasi, "Achieving near-optimum asymptotic efficiency and fading resistance over the time-varying rayleigh-faded cdma channel," *IEEE Transactions on Communications*, 1996.

[8] G. Castellini, F. Conti, E. Re, and L. Pierucci, "A continuously adaptive MLSE receiver for mobile communications: Algorithm and performance," *IEEE Transactions on Communications*, 1997.

[9] C. Georgiades and J. Han, "Sequence estimation in the presence of random parameters via the EM algorithm," *IEEE Transactions on Communications*, 1997.

Multiuser Detection with Random Spreading and Error-Correction Codes: Fundamental Limits

Sergio Verdú, Princeton University
Shlomo Shamai (Shitz), Technion

Abstract

The CDMA channel with randomly and independently chosen spreading sequences accurately models the situation where pseudo-noise sequences span many symbol periods. Its analysis provides a comparison baseline for CDMA channels with carefully designed signature waveforms that span one bit period. We analyze the spectral efficiency (total capacity per chip) as a function of the number of users, spreading gain and signal-to-noise ratio, and we quantify the loss in efficiency relative to an optimally chosen set of signature sequences and to a multiaccess system without spreading. White Gaussian background noise and equal-power synchronous users are assumed. The following receivers are analyzed: a) optimal joint processing, b) single-user matched filtering; c) decorrelation and d) MMSE linear processing.

I Spectral Efficiency

Each signature in a direct-sequence spread-spectrum code division multiple access system can be viewed as a unit-norm vector in an N-dimensional signal space, where N is the spreading gain or number of chips per symbol. In the model considered in this paper, K users linearly modulate their signatures with the outputs of respective autonomous encoders which map information bits to channel symbols. The central question we address is the capacity loss incurred by the imposition of such a structure on the transmitted signals and by the adoption of several popular multiuser detectors. The fundamental figure of merit in a coded spread spectrum system is the *spectral efficiency*, C, defined as the total number of bits per chip that can be transmitted arbitrarily reliably. Since the bandwidth of the CDMA system is (roughly) equal to the reciprocal of the chip duration, the spectral efficiency can be viewed as the bits/s/Hz supported by the system. Note that if the individual code rates (bits/symbol) are identical and denoted by R, then the spectral efficiency is equal to the product:

$$\mathsf{C} = \frac{K}{N} R. \tag{1}$$

In a system where no spreading is imposed, the encoders are able to control the symbols modulating each chip independently. Therefore, assuming chip-synchronism, the Cover-Wyner capacity region of the conventional Gaussian multiaccess channel [5] applies to this case and the spectral efficiency in the absence of spreading is given by

$$\mathsf{C}^* = \frac{1}{2}\log\left(1 + \frac{K}{N}\mathsf{SNR}\right). \tag{2}$$

where, for consistency with the results below, SNR denotes the energy per transmitted N chips divided by the Gaussian noise spectral level, σ^2. This means that the energy per bit divided by $N_0 = 2\sigma^2$ is

$$\frac{E_b}{N_0} = \frac{\mathsf{SNR}}{2R}. \tag{3}$$

Once the spectral efficiency is determined, it is possible to obtain the minimum bandwidth necessary to transmit a predetermined information rate or the maximum information rate that can be supported by a given bandwidth. In order to compare different systems (with possibly different spreading gains and data rates), the spectral efficiency must be given as a function of $\frac{E_b}{N_0}$. According to (1) and (3), if the system operates at full capacity, then SNR in (2) can be substituted by

$$\mathsf{SNR} = \frac{2N}{K}\frac{E_b}{N_0}\mathsf{C}^*$$

so the maximum spectral efficiency $\mathsf{C}^*\left(\frac{E_b}{N_0}\right)$ in the absence of spreading is the solution to

$$\mathsf{C}^*\left(\frac{E_b}{N_0}\right) = \frac{1}{2}\log\left(1 + 2\mathsf{C}^*\left(\frac{E_b}{N_0}\right)\frac{E_b}{N_0}\right) \tag{4}$$

or equivalently

$$\frac{4^{\mathsf{C}^*\left(\frac{E_b}{N_0}\right)} - 1}{2\mathsf{C}^*\left(\frac{E_b}{N_0}\right)} = \frac{E_b}{N_0}.$$

The solution is well-known to satisfy [3]

$$\mathsf{C}^*\left(\frac{E_b}{N_0}\right) > 0 \quad \text{if and only if } \frac{E_b}{N_0} > \log_e 2 = -1.6 \text{ dB},$$

and

$$\lim_{\frac{E_b}{N_0}\to\infty} \frac{\mathsf{C}^*\left(\frac{E_b}{N_0}\right)}{10\log_{10}\frac{E_b}{N_0}} = \frac{\log 10}{20} = 0.166 \text{ bits/dB}. \tag{5}$$

Since (4) does not depend on K, when the transmitted signals are not constrained to the spread-spectrum format, the spectral efficiency is the same as in a single-user system.

The total capacity (sum-rate) of the power-constrained synchronous CDMA channel

$$y(t) = \sum_{k=1}^{K}\sum_{i=1}^{n} A_k b_k[i]\, s_k(t - iT) + n(t)\,,$$

is equal to [25]

$$\frac{1}{2}\log\left(\det\left[\mathbf{I} + \sigma^{-2}\mathbf{ARA}\right]\right).$$

where $\mathbf{A} = \operatorname{diag}\{A_1, \ldots, A_K\}$, and $\mathbf{R}$ is the matrix of normalized crosscorrelations.

If the users have equal power, then $A_k = A$, $\mathsf{SNR} = A^2/\sigma^2$, and the optimum spectral efficiency is equal to

$$\mathsf{C}^{\mathsf{opt}}(\mathsf{SNR}, \mathbf{R}, K, N) = \frac{1}{2N}\log\left(\det\left[\mathbf{I} + \mathsf{SNR}\mathbf{R}\right]\right) \tag{6}$$

For example, in the case of orthogonal sequences the spectral efficiency is equal to

$$\mathsf{C}^{\mathsf{orth}} = \frac{K}{2N}\log(1 + \mathsf{SNR}), \quad \text{if} \quad K \leq N. \tag{7}$$

Substituting

$$\mathsf{SNR} = \mathsf{C}^{\mathsf{orth}} \frac{2N}{K} \frac{E_b}{N_0},$$

we obtain that if $K \leq N$, then

$$\mathsf{C}^{\mathsf{orth}} \left(\frac{K}{N}, \frac{E_b}{N_0} \right) = \frac{K}{N} \mathsf{C}^{*} \left(\frac{E_b}{N_0} \right). \tag{8}$$

The equality of $\mathsf{C}^{\mathsf{orth}}$ and C^{*} for $K = N$ is a consequence of the well-known fact [5] that for equal-rate equal-power users orthogonal multiple access incurs no loss in capacity relative to unconstrained multiple access. It is also known [16] that even if $K > N$, there exist spreading codes (*Welch-Bound-Equality*) which incur no loss in capacity.

Optimal spectral efficiency in non-orthogonal CDMA requires joint processing and decoding of users. As advocated in a number of recent works [17, 6, 18, 1, 23, 10, 26, 4], it is sensible in terms of complexity-performance tradeoff to adopt as a front-end a (soft-output) *multiuser detector* [24] followed by autonomous single-user error-control decoders. In our analysis of spectral efficiency we consider, in addition to optimal decoding, some popular linear multiuser detector front-ends:

- Single-user matched filter.
- Decorrelator.
- Linear Minimum Mean-Square-Error (MMSE).

Our purpose is to evaluate the spectral efficiency of CDMA systems where *signature waveforms are assigned at random.* Denote the unit-norm signature of the kth user by $[c_{k1}, \cdots c_{kN}]$, and assume that $c_{kj} \in \{-1/\sqrt{N}, +1/\sqrt{N}\}$ are chosen equally likely and independently for all (k, j). The rationale for averaging capacity with respect to random signature waveforms is twofold:

- It accurately models CDMA systems (such as IS-95 [20]) where pseudo-noise sequences span many symbol periods.
- The spectral efficiency averaged with respect to the choice of signatures provides a lower bound to the optimum spectral efficiency achievable with a deterministic choice of signature waveforms.

In our asymptotic (in K) analysis we do not just average spectral efficiency with respect to the spreading sequences, but we show convergence of the spectral efficiencies to deterministic quantities.

Most analyses of multiuser detectors have focused on the bit-error-rate of uncoded communication ([24]). The results found on this paper give the best achievable performance with error-control coding assuming random signature waveforms (Figure 1). This serves as a lower bound to the performance achievable through careful design of signatures with low-crosscorrelations. Furthermore this analysis is directly applicable to multiuser detectors operating with spreading codes whose periodicity is much larger than the spreading gain (e.g. [27],[13]).

II Optimal Decoding

Optimal decoding can be performed by a bank of matched filters (which converts the received process to a discrete-time vector process) followed by joint maximum likelihood decoding of the error control code. The formula in [25] for capacity as a function of the signature waveforms was used in [16] to show that if so-called Welch-bound-equality signature waveforms are assigned, then the spectral efficiency of the CDMA system is equal to the case of no-spreading (2); a necessary condition for the existence of such signature waveforms is $K \geq N$. It had been conjectured in [14] that as $K/N \to \infty$ and $\mathsf{SNR} \to \infty$ the loss incurred by a random choice of signatures vanishes. This was verified independently by MonteCarlo simulation in [26] and with an asymptotic $K \geq N \to \infty$ lower bound on the average capacity for random signature waveforms in [7, 8]. We have shown that the gain in optimum spectral efficiency achievable by allocating instantaneous power as a function of the actual spreading waveform employed by other users is negligible. Henceforth, all the results assume that power is not allocated dynamically depending on the realization of $\mathbf{R}$. When $K = N = 2$, we have shown that binary random sequences achieve 75% of the spectral efficiency of orthogonal sequences. Using recent results on the asymptotic spectral distribution of random covariance matrices [2] we have shown the following result (proofs are omitted in this conference version because of space limitations) in the asymptic case of $K \to \infty$ and $K = \beta N$.

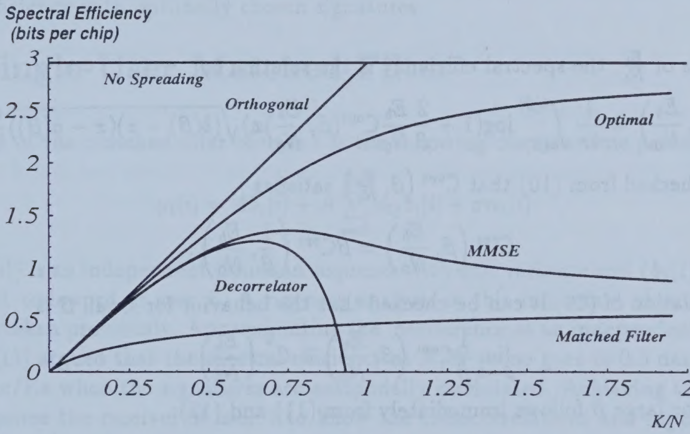


Figure 1: Large–K Spectral Efficiencies for $\frac{E_b}{N_0} = 10$ dB.

Proposition II.1. *Let $\beta > 0$, then the following limit holds almost surely*

$$\lim_{K\to\infty} \mathsf{C}^{\mathsf{opt}}(\mathsf{SNR}, \mathrm{R}, K, K/\beta) = \frac{1}{4\pi}\int_{a(\beta)}^{b(\beta)} \log(1+\mathsf{SNR}x)\sqrt{(b(\beta)-x)(x-a(\beta))}\frac{dx}{x}, \quad (9)$$

where $a(\beta) = (\sqrt{\beta}-1)^2$, *and* $b(\beta) = (\sqrt{\beta}+1)^2$.

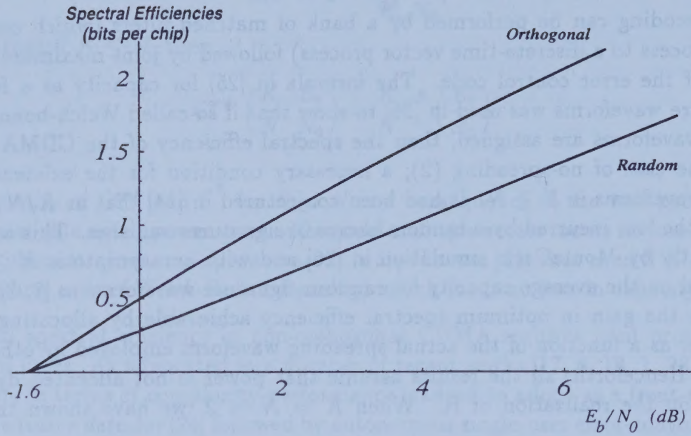


Figure 2: Optimum Spectral Efficiencies with Orthogonal and Random Sequences $K = N \to \infty$.

In terms of $\frac{E_b}{N_0}$, the spectral efficiency is the solution to

$$\mathsf{C}^{\text{opt}}\left(\beta, \frac{E_b}{N_0}\right) = \frac{1}{4\pi}\int_{a(\beta)}^{b(\beta)} \log(1 + \frac{2}{\beta}\frac{E_b}{N_0}\mathsf{C}^{\text{opt}}(\beta, \frac{E_b}{N_0})x)\sqrt{(b(\beta) - x)(x - a(\beta))}\,\frac{dx}{x} \qquad (10)$$

It can be checked from (10) that $\mathsf{C}^{\text{opt}}\left(\beta, \frac{E_b}{N_0}\right)$ satisfies

$$\mathsf{C}^{\text{opt}}\left(\beta, \frac{E_b}{N_0}\right) = \beta\mathsf{C}^{\text{opt}}\left(\frac{1}{\beta}, \frac{E_b}{N_0}\right). \qquad (11)$$

By manipulation of (10) it can be checked that the behavior for small β is

$$\lim_{\beta\to 0}\frac{1}{\beta}\mathsf{C}^{\text{opt}}\left(\beta, \frac{E_b}{N_0}\right) = \mathsf{C}^*\left(\frac{E_b}{N_0}\right). \qquad (12)$$

The limit for large β follows immediately from (11) and (12):

$$\lim_{\beta\to\infty}\mathsf{C}^{\text{opt}}\left(\beta, \frac{E_b}{N_0}\right) = \mathsf{C}^*\left(\frac{E_b}{N_0}\right). \qquad (13)$$

It is straightforward to show that as $\frac{E_b}{N_0} \to \infty$ the slope of the spectral efficiency achievable with random sequences as a function of $\frac{E_b}{N_0}$ coincides with (5) for $\beta \geq 1$. In Figure 2 we have shown $\mathsf{C}^{\text{opt}}\left(1, \frac{E_b}{N_0}\right)$ and $\mathsf{C}^{\text{orth}}\left(1, \frac{E_b}{N_0}\right)$. The slopes of both curves with the logarithm of $\frac{E_b}{N_0}$ are asymptotically equal. However, there is a gap between both curves which can be shown to be half a nat (0.72 bits) asymptotically.

If $\frac{E_b}{N_0}$ is close to its lower limit of $\log_e(2) = -1.6$ dB, then random sequences only achieve 50% of the spectral efficiency of orthogonal sequences. Figure 3 displays the percentage of the spectral efficiency of optimum sequences which is achieved by random sequences, When K is large, the loss in spectral efficiency as a function of E_b/N_0 due to a random choice of sequences (as opposed to optimal) vanishes as $E_b/N_0 \to \infty$ or as $\beta \to \infty$. The maximum loss is 50% and occurs at $K = N$, $E_b/N_0 \downarrow \log_e 2$.

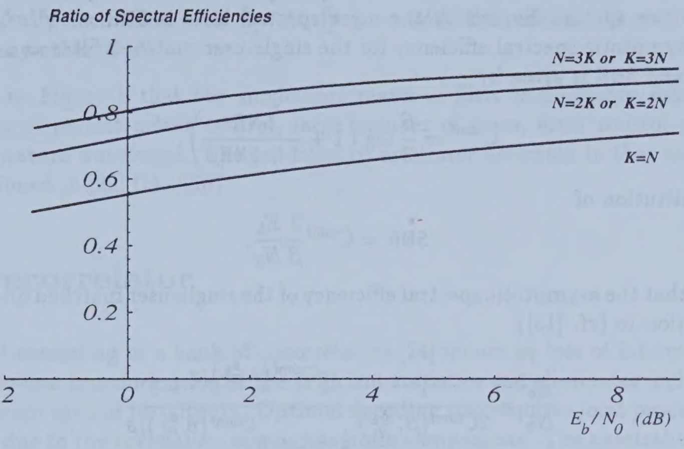


Figure 3: Optimal Processing: Spectral Efficiency with random signatures divided by spectral efficiency with optimally chosen signatures.

III Single-User Matched Filter

The output of the matched filter of user 1 is the following discrete-time process:

$$y_1[i] = Ab_1[i] + A\sum_{j=2}^{K} \rho_{1j} b_j[i] + \sigma n_1[i] \tag{14}$$

where $\{n_1[i]\}$ is an independent Gaussian sequence with unit variance and $\{b_k[1], \ldots, b_k[n]\}$ is the input codeword of user k. A rigorous analysis of this important channel has not been undertaken previously. Approximating the interference as an independent Gaussian sequence, [11] argued that the spectral efficiency as $K/N \to \infty$ goes to 0.5 nat/sec/Hz = 0.72 bit/sec/Hz when the signatures are antipodally modulated. Achieving the capacity of (14) requires the receiver of user 1 to know the crosscorrelations and input distributions of all the interferers. However, the following result (proved using a recent central limit theorem [12]) shows that that information becomes useless as the number of users grows without bound. Furthermore, no averaging of capacity with respect to the random signature sequences is required.

Proposition III.1. *Let* $\mathsf{C}(\rho_{12}, \ldots, \rho_{1K}, p_{b_2}, \ldots, p_{b_K})$ *denote the capacity of the single-user channel (14) where the input codeword* $\{b_1[1], \ldots, b_1[n]\}$ *is constrained to satisfy*

$$\frac{1}{n}\sum_{i=1}^{n} b_1^2[i] \le 1,$$

the random variables $\{b_j[i]\}$ *are independent with distribution* p_{b_k}, $E[b_j^2[i]] = 1$, *and* $n_1[i]$ *is Gaussian with unit variance. Then, as* $\beta N = K \to \infty$

$$\mathsf{C}(\rho_{12}, \ldots, \rho_{1K}, p_{b_2}, \ldots, p_{b_K}) \xrightarrow{a.s.} \frac{1}{2}\log\left(1 + \frac{A^2}{\sigma^2 + \beta A^2}\right). \tag{15}$$

Equation (15) gives the capacity per user and per symbol (N chips). To obtain the spectral efficiency, all we need to do is multiply by K and divide by N. Recalling that the energy per symbol divided by the noise spectral level is $\mathsf{SNR} = A^2/\sigma^2$, we obtain that the asymptotic spectral efficiency for the single-user matched filter as a function of $\beta = K/N$ and SNR is given by

$$\mathsf{C}^{\mathsf{sumf}} = \frac{\beta}{2}\log\left(1 + \frac{\mathsf{SNR}}{1+\mathsf{SNR}\beta}\right).$$

Upon substitution of

$$\mathsf{SNR} = \mathsf{C}^{\mathsf{sumf}}\frac{2}{\beta}\frac{E_b}{N_0},$$

we obtain that the asymptotic spectral efficiency of the single-user matched filter $\mathsf{C}^{\mathsf{sumf}}(\beta, \frac{E_b}{N_0})$, is the solution to (cf. [15])

$$\frac{E_b}{N_0} = \frac{1}{2\mathsf{C}^{\mathsf{sumf}}(\beta, \frac{E_b}{N_0})}\frac{4^{\mathsf{C}^{\mathsf{sumf}}\left(\beta,\frac{E_b}{N_0}\right)/\beta} - 1}{\frac{1}{\beta} + 1 - 4^{\mathsf{C}^{\mathsf{sumf}}\left(\beta,\frac{E_b}{N_0}\right)/\beta}}. \qquad (16)$$

It follows from (16) that $\mathsf{C}^{\mathsf{sumf}}(\beta, \frac{E_b}{N_0})$ is monotonically increasing with β, and

$$\lim_{\beta\to\infty} \mathsf{C}^{\mathsf{sumf}}(\beta, \frac{E_b}{N_0}) = \frac{\log_2 e}{2} - \frac{1}{2}\frac{N_0}{E_b} \qquad \frac{E_b}{N_0} > \log_e 2. \qquad (17)$$

which is seen to converge to one half nat as $\frac{E_b}{N_0} \to \infty$. The use of random signatures as opposed to optimally chosen sequences brings about substantial losses in spectral efficiency for the single-user matched filter, unless $\frac{E_b}{N_0}$ is relatively low and K/N is high.

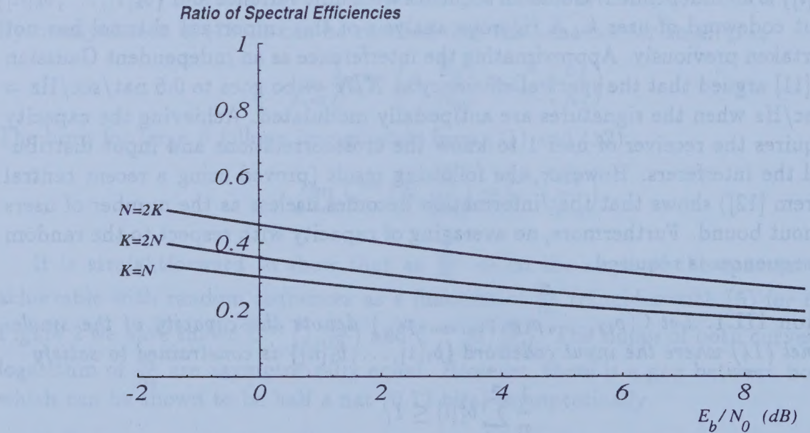


Figure 4: Single-user Matched Filtering: Spectral Efficiency with random signatures divided by spectral efficiency with optimally chosen signatures.

Figure 4 shows the loss due to a random choice of signatures relative to optimally designed sequences. Recall that if $K \leq N$, orthogonal sequences are optimal, and if

$K = mN$, they remain optimal provided each is assigned to m users. In either case, the single-user matched filter is an optimal front-end. The ratio of spectral efficiencies is monotonically decreasing with $\frac{E_b}{N_0}$. At $\frac{E_b}{N_0} \downarrow -1.6$dB, the ratio is $1/3$ at $K = N$, and higher for any other K/N.

We see in Figure 1 that the single-user matched filter is distinctly suboptimal in the scenario of perfect power control, large number of users, error control coding and random signature waveforms. The potential of multiuser detection in that scenario had been questioned in [22] (cf. [23]).

IV Decorrelator

A front-end consisting of a bank of decorrelators [24] incurs no loss of information since it is a one-to-one transformation of the sufficient statistics and eliminates multiaccess interference from each of its outputs. Optimal decoding still requires joint processing of all K outputs due to the correlation among the noise components. The spectral efficiency of this method is maximum. The point of studying capacity with a decorrelating front-end is that it lends itself naturally to a suboptimal approach in which single-user decoding is based on each individual decorrelator (unquantized) output. Since the output of each single-user decorrelator is uncontaminated by multiaccess interference, the analysis of the single-user decorrelator capacity requires the single-user capacity formula evaluated at the decorrelator output signal-to-noise ratio, which is equal to the maximum near-far resistance, $\bar{\eta}$, [24]. Bounds and MonteCarlo simulation of capacity using the decorrelator were given in [26] and [4]. Random sequences with complex-valued chips where sequences are uniformly distributed on the surface of the unit-radius N-dimensional sphere are considered in [15, 19] where an expression is found for the density function of the maximum near-far resistance as a function of K and N, We have shown the following result:

Proposition IV.1. *For $\beta \leq 1$ and binary random spreading, the spectral efficiency of the decorrelator converges in mean-square as $K \to \infty$ to*

$$\lim_{K\to\infty} \mathsf{C}^{\mathsf{deco}} = \frac{\beta}{2} \log\left(1 + \mathsf{SNR}(1-\beta)\right), \tag{18}$$

Upon comparison to (4), we get

$$\mathsf{C}^{\mathsf{deco}}\left(\beta, \frac{E_b}{N_0}\right) = \beta \mathsf{C}^*\left((1-\beta)\frac{E_b}{N_0}\right). \tag{19}$$

Notice that (cf. Figure 1) $\mathsf{C}^{\mathsf{deco}} = 0$ if

$$\frac{E_b}{N_0} \leq \frac{\log_e 2}{1-\beta}.$$

Since the decorrelator is an optimum front-end in the case of orthogonal sequences, the loss due to the use of random sequences is given by (via (8) and (19))

$$\gamma^{\mathsf{deco}}\left(\beta, \frac{E_b}{N_0}\right) \stackrel{\text{def}}{=} \frac{\mathsf{C}^{\mathsf{deco}}\left(\beta, \frac{E_b}{N_0}\right)}{C^{\mathsf{orth}}\left(\beta, \frac{E_b}{N_0}\right)} = \frac{\mathsf{C}^*\left((1-\beta)\frac{E_b}{N_0}\right)}{C^*\left(\frac{E_b}{N_0}\right)} \quad \text{if } \beta \leq 1$$

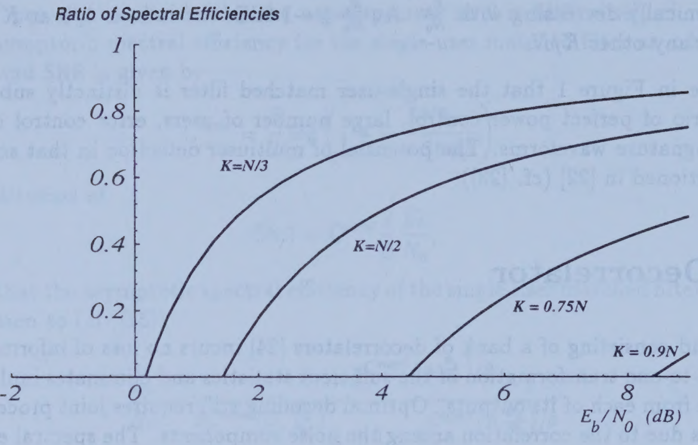


Figure 5: Decorrelator: Spectral Efficiency with random signatures divided by spectral efficiency with optimally chosen signatures.

which is shown in Figure 5. Comparing this figure to Figure 3 we can see that for high $\frac{E_b}{N_0}$ and low β the decorrelator achieves near optimal spectral efficiency (see also Figure 1). As should be expected and in contrast to the single-user matched filter, the suboptimality of the random choice decreases with $\frac{E_b}{N_0}$.

V MMSE Receiver

The linear MMSE receiver [24] offers a compromise between the multiaccess suppression capabilities of the decorrelator and the optimal background-noise combating capabilities of the single-user matched filter. Unlike the decorrelator, the MMSE filter is well-defined regardless of whether K is smaller or larger than N. As in the case of the decorrelator, we are interested in the spectral efficiency of the MMSE linear transformation followed by single-user decoding. When the channel symbols are binary and binary decisions are made at the output of the linear transformation, [21] shows that the spectral efficiency (of both the matched filter and of the MMSE transformation) tends to 0.46 bits/sec/Hz as $K/N \to \infty$ in the synchronous case and to 0.69 bits/sec/Hz in the asynchronous case. MonteCarlo simulation of the expected capacity with binary sequences and (nonbinary) power-constrained codewords and the MMSE transformation was given in [26]. We have obtained the following result:

Proposition V.1. *For $\beta > 0$ and binary random spreading, the spectral efficiency of the MMSE receiver converges in mean-square as $K \to \infty$ to*

$$\lim_{K\to\infty} \mathsf{C}^{\mathsf{mmse}} = \frac{\beta}{2} \log\left(1 + \mathsf{SNR} - \frac{1}{4}\mathcal{F}(\mathsf{SNR}, \beta)\right)$$

where

$$\mathcal{F}(x,z) \stackrel{\text{def}}{=} \left(\sqrt{x(1+\sqrt{z})^2+1} - \sqrt{x(1-\sqrt{z})^2+1}\right)^2.$$

Let us study the behavior of the spectral efficiency of the MMSE receiver for asymptotically large $\frac{E_b}{N_0}$. For $0 < \beta < 1$, the MMSE and decorrelator spectral efficiencies coincide (cf.Figure 7) :

$$\lim_{\frac{E_b}{N_0}\to\infty} \mathsf{C}^{\mathsf{mmse}}\left(\beta, \frac{E_b}{N_0}\right) - \mathsf{C}^{\mathsf{deco}}\left(\beta, \frac{E_b}{N_0}\right) = 0.$$

If $\beta = 1$, then it can be shown that

$$\lim_{\frac{E_b}{N_0}\to\infty} \frac{\mathsf{C}^{\mathsf{mmse}}\left(1, \frac{E_b}{N_0}\right)}{\mathsf{C}^{*}\left(\frac{1}{2}\frac{E_b}{N_0}\right)} = \frac{1}{2}.$$

If $\beta > 1$, then it can be shown that

$$\lim_{\frac{E_b}{N_0}\to\infty} \mathsf{C}^{\mathsf{mmse}}\left(\beta, \frac{E_b}{N_0}\right) = \frac{\beta}{2}\log\frac{\beta}{\beta-1}.$$

The asymptotic behavior of the spectral efficiency of the MMSE receiver with $\beta \to \infty$ is identical to that of the single-user matched filter:

$$\lim_{\beta\to\infty} \mathsf{C}^{\mathsf{mmse}}\left(\beta, \frac{E_b}{N_0}\right) = \frac{\log_2 e}{2} - \frac{1}{2}\frac{N_0}{E_b} \qquad \frac{E_b}{N_0} > \log_e 2. \tag{20}$$

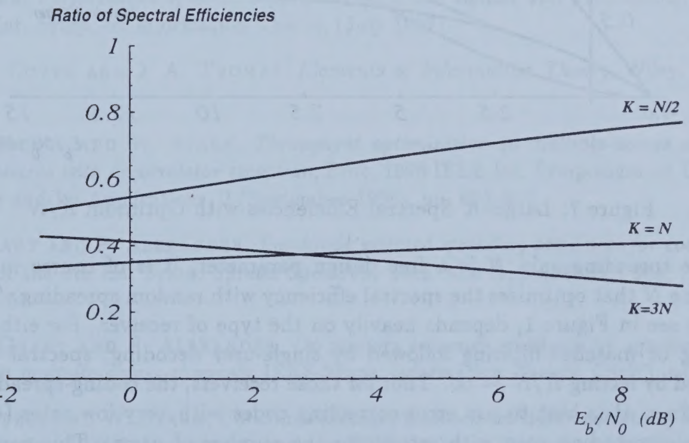


Figure 6: MMSE Receiver: Spectral Efficiency with random signatures divided by spectral efficiency with optimally chosen signatures.

Comparing Figure 6 to Figure 3 we see that a random choice of spreading codes (as opposed to optimal) brings about substantial losses for $K = 3N$. For $K = N$, and in the range of $\frac{E_b}{N_0}$ considered in Figure 6 a random choice achieves around 40% of the spectral efficiency achieved by orthogonal sequences. As the spreading gain N increases, the MMSE detector loss is more important at low $\frac{E_b}{N_0}$ and approaches that in Figures 3 and 5 for large $\frac{E_b}{N_0}$. The deleterious effect of low $\frac{E_b}{N_0}$ on the decorrelator (Figure 5) is not suffered by the MMSE receiver. Relative to Figure 4, we see that at low $\frac{E_b}{N_0}$ the MMSE and single-user matched filter behave similarly; at high $\frac{E_b}{N_0}$ the comparison depends heavily on K/N.

When K/N is large, linear multiuser detection is distinctly suboptimal (Figure 1); however, for lightly loaded systems, the performance improvement afforded by nonlinear techniques is quite small.

VI Optimum Coding-Spreading Tradeoff

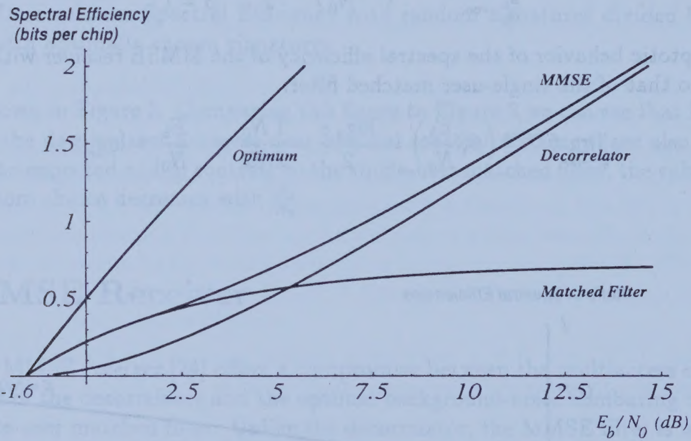


Figure 7: Large–K Spectral Efficiencies with Optimum K/N .

When the spreading gain N is a free design parameter, it is of course interesting to analyze the N that optimizes the spectral efficiency with random spreading. The answer, as we can see in Figure 1, depends heavily on the type of receiver. For either optimum processing or matched-filtering followed by single-user decoding, spectral efficiency is maximized by letting $K/N \to \infty$. Thus for those receivers, the coding-spreading tradeoff favors coding: it is best to use error-correcting codes with very low rates (cf. (1)) and a negligible spreading gain with respect to the number of users. This conclusion was known to hold for the single-user matched filter [11]. Note, however, that the behavior of optimum processing and the conventional single-user matched filter at $K/N \to \infty$ are quite different: the optimal spectral efficiency grows without bound with $\frac{E_b}{N_0}$, whereas the matched filter efficiency approaches 0.72 bit/sec/Hz monotonically as $\frac{E_b}{N_0} \to \infty$.

For large K, the optimum choice of K/N for the decorrelator ranges from 0 for $\frac{E_b}{N_0} \downarrow$

$-1.6\ dB$ to 1 for $\frac{E_b}{N_0} \to \infty$. The optimum coding-spreading tradeoff of the decorrelator dictates using codes whose rates (bits/symbol) lie between 0 ($\frac{E_b}{N_0} \downarrow -1.6\ dB$) to $\mathsf{C}^{\mathsf{deco}}$ ($\frac{E_b}{N_0} \to \infty$). With an optimum choice of spreading gain, the decorrelator spectral efficiency with random signature waveforms is better than that of the single-user matched filter for $\frac{E_b}{N_0} > 5.2$ dB (Figure 7). Unlike the single-user matched filter, the spectral efficiency of the decorrelator grows without bound as $\frac{E_b}{N_0} \to \infty$.

As far as the optimum coding-spreading tradeoff for the MMSE receiver, for low $\frac{E_b}{N_0}$ it favors making K/N very large in which case the MMSE receiver achieves essentially the same spectral efficiency as the single-user matched filter (Figure 7). The optimum K/N reaches 1 at $\frac{E_b}{N_0} = 4$ dB, and a minimum of 0.75 at $\frac{E_b}{N_0} = 10$ dB.

Acknowledgement: This work was supported by the US-Israel Binational Science Foundation and the U.S. Army Research Office under Grant DAAH04-96-1-0379. We are grateful to Amir Dembo, Michael Honig and Emre Telatar for helpful discussions.

References

[1] P. Alexander, L. Rasmussen, and C. Schlegel, *A linear receiver for coded multiuser CDMA*, IEEE Trans. on Communications, vol. 45 no. 5 (May 1997), pp. 605–610.

[2] Z. D. Bai and Y. Q. Yin, *Limit of the smallest eigenvalue of a large dimensional sample covariance matrix*, Annals of Probability, 21 (1993), pp. 1275–1294.

[3] R. Blahut, *Digital Transmission of Information*, Addison-Wesley, Reading, Mass, 1990.

[4] A. Burr, *Performance of linear separation of CDMA signals with FEC coding*, Proc. 1997 IEEE Int. Symp. on Information Theory, (July 1997).

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[6] D. Goeckel and W. Stark, *Throughput optimization in multiple-access communication systems with decorrelator reception*, Proc. 1996 IEEE Int. Symposium on Information Theory and Its Applications, 2 (September 1996), pp. 653–656.

[7] A. Grant and P. Alexander, *Randomly selected spreading sequences for coded CDMA*, Proc. of the 5th Int. Symp. Spread Spectrum Tech. and Appl.(ISSSTA'96), (Sep. 1996), pp. 54–57.

[8] A. J. Grant and P. Alexander, *On random sequence multisets for synchronous code-division multiple-access channels*, IEEE Trans. Information Theory, (1997). submitted.

[9] A. Hafeez and W. Stark, *Combined decision-feedback multiuser detection/soft-decision decoding for CDMA channels*, to appear in IEEE Vehicular Technology Conference, (1996).

[10] ——, *A family of soft-output multiuser demodulators for coded asynchronous CDMA channels*, Proc. Conf. on Interference Rejection and Signal Separation in Wireless Communications, (Mar. 1997).

[11] J. Y. N. Hui, *Throughout analysis for code division multiple accessing of the spread spectrum channel*, IEEE J. Selec. Areas in Commun., SAC-2 (July 1984), pp. 482–486.

[12] O. JOHNSON, *Entropy inequalities and the central limit theorem*, tech. rep., Statistical Laboratory, Cambridge University, August 1997.

[13] N. MANDAYAM AND S. VERDÚ, *Analysis of an approximate decorrelating detector*, Wireless Personal Communications, (1997). to appear.

[14] J. L. MASSEY, *Is the choice of spreading sequences important?*, Proceedings IEEE International Symposium on Spread Spectrum Techniques Applications, (ISSSTA '96), (Sep. 22-25, 1996).

[15] R. R. MÜLLER, P. SCHRAMM, AND J. B. HUBER, *Spectral efficiency of CDMA systems with linear interference suppression*, IEEE Workshop Communications Engineering, (1997).

[16] M. RUPF AND J. L. MASSEY, *Optimum sequences multisets for synchronous code-division multiple-access channels*, IEEE Trans. Inform. Theory, 40 (July 1994), pp. 1261–1266.

[17] M. RUPF, F. TARKOY, AND J. L. MASSEY, *User-separating demodulation for code-division multiple-access systems*, IEEE J. Selec. Areas in Commun., 2 (June 1994), pp. 786–795.

[18] C. SCHLEGEL, S. ROY, P. ALEXANDER, AND Z. XIANG, *Multi-user projection receivers*, IEEE J. Selected Areas on Communications, 14 (Oct. 1996), pp. 1610–1618.

[19] P. SCHRAMM, R. MUELLER, AND J. HUBER, *Spectral efficiency of multiuser systems based on CDMA with linear MMSE interference suppression*, Proc. IEEE International Symposium on Information Theory, (June-July 1997), p. 357.

[20] TELECOMMUNICATIONS INDUSTRY ASSOCIATION, TIA/EIA, *Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System IS-95A*, Washington, DC, 1995.

[21] V. VEERAVALLI AND B. AAZHANG, *On the coding-spreading tradeoff in CDMA systems*, 1996 Conf. Inform. Science and Syst., (March 1996), pp. 1136–1146.

[22] S. VEMBU AND A. VITERBI, *Two different philosophies in CDMA-a comparison*, Proc. IEEE Vehicular Technology Conference, (Apr.-May 1996), pp. 869–873.

[23] S. VERDÚ, *Demodulation in the presence of multiaccess interference: progress and misconceptions*, in Intelligent Methods in Signal Processing and Communications, D. Docampo, A. Figueiras, and F. Perez, eds., Birkhauser, 1997, ch. 2.

[24] ——, *Multiuser Detection*, Cambridge University Press, New York, 1998.

[25] ——, *Capacity region of Gaussian CDMA channels: The symbol-synchronous case*, Proceedings of the Twenty-Fourth Annual Allerton Conf. Communication, Control, and Computing, (Oct. 1986), pp. 1025–1039.

[26] B. VOJCIC, *Information theoretic aspects of multiuser detection*, Proc. IRSS'97 Interference Rejection and Signal Separation in Wireless Communications, (Mar. 1997).

[27] B. ZAIDEL, S. S. (SHITZ), AND H. MESSER, *Performance of linear MMSE front end combined with standard IS-95 uplink*, Wireless Networks (Special Issue: Multiuser Detection in Wireless Communications), (1997).

Distributed Estimation Techniques for Wireless Sensor Networks

Haralabos C. Papadopoulos and Gregory W. Wornell
Research Laboratory of Electronics
Massachusetts Institute of Technology

Abstract

We focus our attention on a problem of signal estimation via quantized measurements that arises in a number of wireless sensor networks. A central site estimates an information-bearing signal from low-bandwidth information received from remote sensors, and may or may not broadcast feedback information to the sensors. We examine the use of an additive control input before signal quantization as a low-complexity encoding strategy that can potentially enhance the estimation quality. We develop efficient encoding strategies and associated estimators in the context of pseudo-noise control inputs, fully known control waveforms, and control inputs that exploit past quantized output values, which provide increasingly better performance over the associated control-free system.

1 Introduction

Quantizers are important elements in a wide range of signal processing and communications applications. For instance, they are often used to provide efficiently-encoded signal representations, such as in the context of several existing and future wireless sensor networks, where remote sensors communicate digitally-encoded signals to a central site via a bandwidth-constraint channel.

An important instance of such a remote sensing scenario involves a central site estimating a remotely-acquired analog signal from an efficiently-encoded digital description. Depending upon bandwidth availability, the central site may or may not broadcast information back to these sensors. We consider a low-complexity encoding strategy according to which a control input is added to the information-bearing signal before digitization at each remote sensor. The block diagram corresponding to a single sensor in the context of such a remote sensing estimation environment is shown in Fig. 1, where $A[n]$ denotes the slowly-varying information-bearing signal, $v[n]$ represents sensor noise, $w[n]$ is a control input, and $y[n]$ denotes the quantized signal that is sent to the central site.

We focus on the special case of the estimation problem depicted in Fig. 1 in which $A[n] = A$, *i.e.*, we examine the problem of estimating a noise-corrupted unknown parameter A via quantized observations. This case reveals some of the key features of estimation from quantized observations obtained via a system comprising a control input and a quantizer. Extensions to slowly-varying input signals are developed in [1].

This work has been supported in part by DARPA monitored by ONR under Contract No. N00014-93-1-0686, AFOSR under Grant No. F49620-96-1-0072, and ARL under Cooperative Agreement No. DAAL01-96-2-0002.

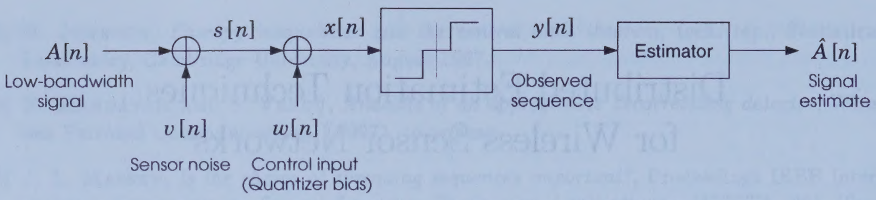


Figure 1: Signal estimation via quantized observations, in the context of low-complexity encoding methods that are based on an additive control input.

Several basic variations of the estimation problem in Fig. 1 can arise in practice, which differ in the control information that is available for estimation and the associated freedom in the control input selection. For pseudo-noise control inputs whose statistical characterization alone is exploited at the receiver, we show that there is an optimal power level for minimizing the mean-square estimation error (MSE). The existence of a non-zero optimal pseudo-noise power level reveals strong connections to the phenomenon of stochastic resonance, which is encountered in a number of physical nonlinear systems where thresholding occurs [2]. Performance can be further enhanced if detailed knowledge of the control waveform is exploited at the receiver. In this scenario, we develop methods for judiciously selecting the control input from a suitable class of periodic waveforms for any given system. Finally, if feedback from the quantized output to the control input is available, we show that, when combined with suitably designed receivers, these signal quantizers come within a small loss of the quantizer-free performance. Although the use of feedback from the central site to the sensor is not required in a single-sensor system since the sensor already knows $y[n]$, it is appealing in wireless sensor networks, where the central site may broadcast control information based on all the available observations.

2 System Model

As outlined in Sec. 1, we consider the problem of estimating an unknown parameter A from observation of

$$y[n] = F(A + v[n] + w[n]) \qquad n = 1, 2, \cdots, N\,, \tag{1}$$

where the sensor noise $v[n]$ is an independent identically-distributed (IID) process, $w[n]$ is a control input, and the function $F(\cdot)$ is an M-level quantizer, with the quantized output $y[n]$ taking M distinct values $Y_1, Y_2, \cdots, Y_M$, *i.e.*,

$$F(x) = \begin{cases} Y_i & \text{if } X_{i-1} < x \le X_i, \quad i = 1, 2, \cdots, M-1 \\ Y_M & \text{otherwise} \end{cases}\,, \tag{2}$$

where $X_0 = -\infty$. For convenience, we also define the intermediate sequence

$$s[n] \triangleq A + v[n] = A + \sigma_v\, \tilde{v}[n]\,. \tag{3}$$

We restrict our attention to processes $v[n]$ and $w[n]$ (in the case that $w[n]$ is viewed as a pseudo-noise sequence for estimation) that are Gaussian. In [3], we extend our results to IID noise processes with PDFs that are non-zero and smooth (C^1) almost everywhere.

3 Performance Limits

In this section we quantify the performance degradation that results from estimating A based on observation of $y[n]$ instead of $s[n]$. For convenience, we introduce the following concept of *information loss*, which we will use as a figure of merit to design quantizer systems and evaluate the associated estimators. We define the information loss for a quantizer system as the ratio of the Cramér-Rao bounds for unbiased estimates of the parameter A obtained via $y[n]$ and $s[n]$, respectively, *i.e.*,

$$\mathcal{L}(A) \triangleq \frac{\mathcal{B}\left(A;\ \mathbf{y}^N\right)}{\mathcal{B}\left(A;\ \mathbf{s}^N\right)}\,, \tag{4}$$

where $\mathcal{B}\left(A;\ \mathbf{y}^N\right)$ is the Cramér-Rao bound for unbiased estimation of A from

$$\mathbf{y}^N \triangleq \begin{bmatrix} y[1] & y[2] & \cdots & y[N] \end{bmatrix}^T\,, \tag{5}$$

and where $\mathcal{B}\left(A;\ \mathbf{s}^N\right)$ and $\mathbf{s}^N$ are defined similarly. We often consider the information loss (4) in dB; it represents the additional MSE in dB that arises from observing $y[n]$ instead of $s[n]$ in the context of efficient estimation of A. From this perspective, better systems achieve smaller information loss over the range of parameter values of interest.

Taking into account the inherent dynamic range limitations of these quantizers, we assume that the unknown parameter takes values in the range $(-\Delta,\ \Delta)$, with Δ assumed to be known. The degradation of the estimation quality can be conveniently characterized in terms of the ratio $\chi = \Delta/\sigma_v$, which we may view as a measure of peak signal-to-noise ratio (SNR).

Worst-case performance is used to characterize the overall system. Accordingly, we define the worst-case information loss via

$$\mathcal{L}_{\max}(\Delta) \triangleq \sup_{|A|<\Delta} \mathcal{L}(A)\,. \tag{6}$$

The worst-case information loss is a function of other system parameters, such as σ_v and $F(\cdot)$, the dependence on which is suppressed for convenience in the above definition.

3.1 Pseudo-noise Control Inputs

We next consider control inputs $w[n]$ that correspond to sample paths of an IID Gaussian process with variance σ_v^2, independent of $v[n]$, and determine the performance limits in estimating A from $\mathbf{y}^N$ by exploiting the statistical characterization of $w[n]$ at the receiver. In particular, we wish to select σ_w so as to minimize the associated worst-case information loss.

Since $\alpha[n] = v[n] + w[n]$ is an IID sequence, the Cramér-Rao bound for all unbiased estimates of the parameter A based on observation of the vector $\mathbf{y}^N$ satisfies

$$\mathcal{B}\left(A;\ \mathbf{y}^N\right) = \frac{1}{N}\,\mathcal{B}\left(A;\ y\right)\,, \tag{7}$$

where $\mathcal{B}\left(A;\ y\right)$ is the Cramér-Rao bound for estimating A based on a single sample of the IID sequence $y[n]$. By taking the second partial derivative of the associated log-likelihood

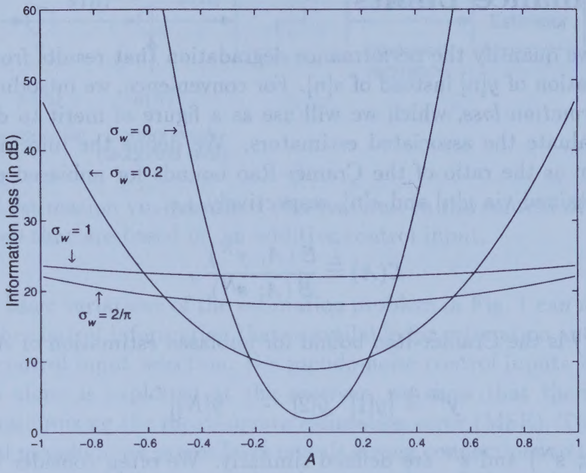


Figure 2: Information loss for a system comprising a two-level quantizer and an IID Gaussian pseudo-noise control input, for various pseudo-noise power levels σ_w. The sensor noise is IID Gaussian with variance $\sigma_v^2 = 0.01$.

function with respect to A followed by an expectation, and by letting $X_M = \infty$, we obtain

$$\mathcal{B}\left(A;\ y\right) = \sigma_\alpha^2 \left(\sum_{i=1}^{M} \frac{\left[f\left(\frac{X_{i-1}-A}{\sigma_\alpha}\right) - f\left(\frac{X_i-A}{\sigma_\alpha}\right)\right]^2}{Q\left(\frac{X_{i-1}-A}{\sigma_\alpha}\right) - Q\left(\frac{X_i-A}{\sigma_\alpha}\right)} \right)^{-1}. \tag{8}$$

where $\sigma_\alpha = \sqrt{\sigma_w^2 + \sigma_v^2}$, $f\left(x\right) = \exp\left(-x^2/2\right)/\sqrt{2\,\pi}$, and $Q\left(x\right) = \int_x^\infty f\left(t\right)\,dt$. For the system corresponding to the symmetric two-level quantizer ($M = 2$, $X_1 = 0$), *i.e.*, $F(x) = \mathrm{sgn}\ x$, the Cramér-Rao bound (8) reduces to

$$\mathcal{B}\left(A;\ y\right) = 2\,\pi\,\sigma_\alpha^2\, Q\left(\frac{A}{\sigma_\alpha}\right) Q\left(-\frac{A}{\sigma_\alpha}\right) \exp\left(\frac{A^2}{\sigma_\alpha^2}\right). \tag{9}$$

Fig. 2 depicts the associated information loss (4) as a function of A for $\Delta = 1$, $\sigma_v = 0.1$ and various σ_w levels. In this Gaussian scenario, for any pseudo-noise power level σ_w the information loss is an increasing function of $|A|$. More important, as the figure reveals, the use of a pseudo-noise control input ($\sigma_w \neq 0$) can be very advantageous in terms of reducing the associated worst-case information loss. In particular, the optimal pseudo-noise power level is given by [3]

$$\sigma_w^{\mathrm{opt}} = \begin{cases} \sqrt{\left(\sigma_\alpha^{\mathrm{opt}}\right)^2 - \sigma_v^2} & \sigma_v < \sigma_\alpha^{\mathrm{opt}} \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where $\sigma_\alpha^{\mathrm{opt}}/\Delta \approx 2/\pi$.

The information loss associated with the optimal pseudo-noise level corresponds to the best achievable performance by a particular family of pseudo-noise sources—in this particular example the family of zero-mean Gaussian distributions. For the optimal choice of σ_w, the worst-case information loss can be completely characterized by means of peak SNR $\chi = \Delta/\sigma_v$. In particular, by using (10) with (9) in (4) we obtain the optimal worst-case information loss for the Gaussian scenario with pseudo-noise control, namely,

$$\mathcal{L}_{\max}^{\text{pn}}(\chi) = \begin{cases} 2\pi\, Q(\chi)\, Q(-\chi)\, e^{\chi^2} & 0 \le \chi \le \frac{\pi}{2} \\ \frac{8}{\pi}\, Q\left(\frac{\pi}{2}\right) Q\left(-\frac{\pi}{2}\right) e^{\frac{\pi^2}{4}}\, \chi^2 & \chi \ge \frac{\pi}{2} \end{cases}, \tag{11}$$

where we explicitly indicate that the worst-case information loss is a function of χ.

As (11) reveals, for parameter estimation in Gaussian noise via a two-level quantizer system, the worst-case information loss can be made to grow as χ^2 by judicious selection of a Gaussian pseudo-noise control input. For comparison, the worst-case information loss in the absence of control input grows exponentially with peak SNR. In particular, by substituting for $\mathcal{B}(A;\ y)$ in (6) the expression in (9) with α replaced by v, we obtain

$$\mathcal{L}_{\max}^{\text{free}}(\chi) = 2\pi\, Q(\chi)\, Q(-\chi)\, e^{\chi^2}, \tag{12}$$

which is proportional to $\exp(\chi^2/2)$ for large χ. The results in (11)–(12) extend to quantizers with $M > 2$, *i.e.*, the worst-case information loss grows as $\exp(\chi^2/2)$ for control-free systems, while it can be made to grow as χ^2 for appropriately chosen Gaussian pseudo-noise control inputs [1].

3.2 Known Control Inputs

When the estimator is able to exploit detailed knowledge of the control waveform, improved performance can be obtained. The Cramér-Rao bound for all unbiased estimates of the parameter A given the first N samples of the known control input $w[n]$ and the corresponding observations $y[n]$ will be denoted by $\mathcal{B}\left(A;\ \mathbf{y}^N,\ \mathbf{w}^N\right)$, and satisfies

$$\mathcal{B}\left(A;\ \mathbf{y}^N,\ \mathbf{w}^N\right) = \left[\sum_{n=1}^{N} \left[\mathcal{B}\left(A + w[n];\ y\right)\right]^{-1}\right]^{-1}, \tag{13}$$

where $\mathcal{B}(A;\ y)$ is given by (8), with α replaced by v. Examination of (13) reveals that the corresponding worst-case information loss is a function of the control waveform $\mathbf{w}^N$. Furthermore, for any known control waveform selection strategy, the worst-case information loss grows at least as fast as χ for any $M \ge 2$ [3].

Classes of periodic waveforms parameterized by the period K are appealing candidates for known control inputs, since the are easy to construct and can be chosen so that the worst-case information loss grows at the minimum possible rate. For any class of K-periodic waveforms, optimal selection of the control input reduces to selecting the period K that minimizes the associated worst-case information loss, or equivalently, using (13)

$$K_{\text{opt}}(\Delta, \sigma_v) \triangleq \arg\min_{K} \sup_{A\,\in\,(-\Delta,\,\Delta)} \frac{K}{\sum_{n=1}^{K} \left[\mathcal{B}\left(A + w[n];\ y\right)\right]^{-1}}, \tag{14}$$

where $\mathcal{B}(A;\ y)$ is given by (8) with α replaced by v.

For the system with $M = 2$ in Gaussian sensor noise, the information loss is largest for parameter values that are furthest from the quantizer threshold, as $\mathcal{B}(A;\ y)$ in (9) is an increasing function of $|A|$. Since our objective is to optimize over the worst-case performance, a potentially appealing strategy is to select $w[n]$ so as to minimize the largest distance between any parameter value A in $(-\Delta,\ \Delta)$ and the closest effective quantizer threshold. For this reason we consider K-periodic sawtooth control inputs

$$w[n] = \delta_w \left(-\frac{K-1}{2} + n \bmod K\right) , \tag{15}$$

where $\delta_w = 2\,\Delta/(K-1)$. The net effect of the periodic control input (15) and a symmetric two-level quantizer $F(\cdot)$ is equivalent to a two-level quantizer with a periodically time-varying threshold; it is important to observe that the time-varying quantizer threshold comes within at least $\delta_w/2$ of any possible parameter value once every K samples. In this case, the optimal period K_{opt} from (14) is only a function of peak SNR χ. Let $d(\chi;K)$ denote the normalized spacing between successive samples of $w[n]$ in (15), *i.e.*, $d(\chi;K) = \delta_w/\sigma_v$. The optimal normalized spacing $d_{\text{opt}}(\chi)$, is the normalized spacing associated with the optimal period $K_{\text{opt}}(\chi)$ from (14), *i.e.*, $d_{\text{opt}}(\chi) \triangleq d(\chi; K_{\text{opt}}(\chi))$. In this Gaussian scenario the optimal normalized spacing approaches [3]

$$d_\infty \triangleq \lim_{\chi\to\infty} d_{\text{opt}}(\chi) \approx 2.5851 , \tag{16}$$

while the corresponding worst-case information loss is well approximated by

$$\mathcal{L}^{\text{per}}_{\text{max}}(\chi) \approx 1.4754 \left(\frac{2\chi}{d_\infty} + 1\right) \tag{17}$$

for large χ. Thus, if we select $w[n]$ as in (15) with $K = \lceil 2\chi/d_\infty + 1\rceil$, the worst-case information loss is approximately given by (17) and grows linearly with χ. In other words, this control input selection scheme achieves the optimal growth rate. K-periodic sawtooth waveforms similar to (15) can be used as control inputs in quantizer systems with $M > 2$, and can be chosen so that the worst-case information loss grows linearly with χ. In this case, the existence of multiple quantizer thresholds allows for reduction of the dynamic range that the sawtooth waveforms $w[n]$ need to span [3].

3.3 Control Inputs in the Presence of Feedback

We next consider the scenario where, in addition to knowing the control waveform, the estimator has the option of using feedback from past output observations in the selection of future control input values. Specifically, we consider the problem of estimating the unknown parameter A from observation of (1), where $F(\cdot)$ is given by (2) and where the control input $w[n]$ at time n is a function of all past quantized output values.

Performance limits for systems with feedback can be developed by examining the expression (13) for the Cramér-Rao bound of A from $\mathbf{y}^N$ for a selected control input vector $\mathbf{w}^N$. Examination of (13) reveals that for any A in $(-\Delta,\ \Delta)$ we can obtain information loss equal to $\mathcal{L}(A_\circ)$ by selecting $w[n] = A_\circ - A$. In particular, if there exists a parameter value A_* for which $\mathcal{B}(A;\ y) \geq \mathcal{B}(A_*;\ y)$ for all A in $(-\infty,\ \infty)$ and where $\mathcal{B}(A;\ y)$ is given by (8) with α replaced by v, then using (13) we obtain

$$\mathcal{B}\left(A;\ \mathbf{y}^N,\ \mathbf{w}^N\right) \geq \mathcal{B}\left(A_*;\ y\right)/N , \tag{18}$$

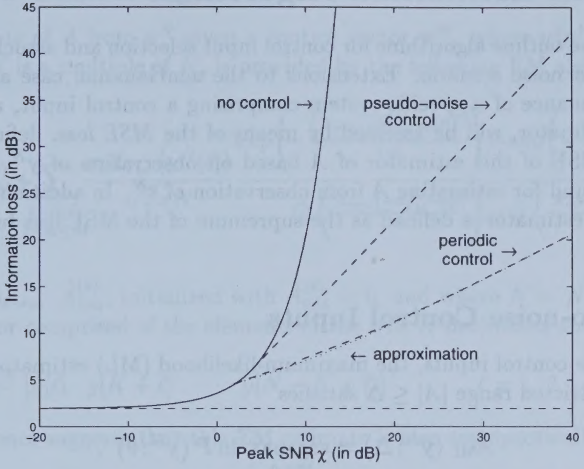


Figure 3: Worst-Case information loss over $|A| < \Delta$ for a two-level quantizer in zero-mean IID Gaussian noise of variance σ_v^2, with no control input (solid curve), pseudo-noise control inputs (upper dashed curve), and known periodic control waveforms (middle dashed curve). The dotted curve depicts approximation (17). The lower dashed line depicts the minimum possible information loss (≈ 2 dB) for any control input scheme.

with equality achieved for $w[n] = A_* - A$ for $n = 1, 2, \cdots, N$. This control input sequence results in

$$\mathcal{L}\left(A; \mathbf{w}^N\right) \geq \mathcal{L}(A; A_* - A) = \mathcal{L}(A_*) \ , \tag{19}$$

where $\mathcal{L}(A)$ is given by (4), and where $\mathcal{B}\left(A;\ y\right)$ is given by (8) with α replaced by v. For the symmetric two-level quantizer in Gaussian sensor noise, use of (9) for $\sigma_\alpha = \sigma_v$ in (6) reveals that $A_* = 0$. From (19), the information loss is lower bounded by

$$\mathcal{L}\left(A; \mathbf{w}^N\right) \geq \mathcal{L}(0) = \frac{\pi}{2} \ , \tag{20}$$

which corresponds to a 2 dB information loss.

Fig. 3 depicts the worst-case information loss for the system corresponding to $M = 2$ in the context of Gaussian sensor noise and the various control input scenarios that we have examined. As the figure reflects, the performance of the control-free system (solid curve) degrades rapidly as χ increases. The benefits of pseudo-noise control inputs (upper dashed curve) at high peak SNR are clearly evident, and known periodic control inputs provide additional performance benefits (middle dashed curve) over pseudo-noise control inputs. In particular, the associated worst-case information loss increases linearly with χ as the accurate approximation (17) reveals. Finally, in the presence of feedback from the quantized output to the control input, the performance is lower bounded by the minimum possible information loss of 2 dB, which is independent of χ. In Sec. 4 we present control selection strategies and associated estimators that meet all these bounds.

4 Efficient Estimation Algorithms

In this section we outline algorithms for control input selection and associated estimators for the Gaussian noise scenario. Extensions to the nonGaussian case are developed in [3]. The performance of a specific system comprising a control input, a quantizer and a particular estimator, will be assessed by means of the *MSE loss*, defined as the ratio of the actual MSE of this estimator of A based on observation of $\mathbf{y}^N$, divided by the Cramér-Rao bound for estimating A from observation of $\mathbf{s}^N$. In addition, the worst-case MSE loss of an estimator is defined as the supremum of the MSE loss function over the range $|A| < \Delta$.

4.1 Pseudo-noise Control Inputs

For pseudo-noise control inputs, the maximum-likelihood (ML) estimator of A based on $\mathbf{y}^N$ over the restricted range $|A| \leq \Delta$ satisfies

$$\hat{A}_{\mathrm{ML}}\left(\mathbf{y}^N;\Delta\right) = \underset{|\theta|\leq\Delta}{\arg\max}\ \ln P\left(\mathbf{y}^N;\theta\right)\ , \tag{21}$$

where $\ln P\left(\mathbf{y}^N;\theta\right)$ is the associated log-likelihood function.

If $F(\cdot)$ is the symmetric two-level quantizer and since $v[n]+w[n]$ is Gaussian, the ML estimate (21) can be found in closed form, by setting to zero the partial derivative of the log-likelihood function with respect to A, *viz.*,

$$\hat{A}_{\mathrm{ML}}\left(\mathbf{y}^N;\Delta\right) = -\mathcal{I}_\Delta\left(\sqrt{\sigma_v^2+\sigma_w^2}\ Q^{-1}\left(\frac{\mathcal{K}_1\left(\mathbf{y}^N\right)}{N}\right)\right)\ , \tag{22}$$

where $Q^{-1}(\cdot)$ is the inverse of $Q(\cdot)$, $\mathcal{K}_{Y_i}\left(\mathbf{y}^N\right)$ denotes the number of elements in $\mathbf{y}^N$ that are equal to Y_i, and $\mathcal{I}_\Delta(\cdot)$ is the following piecewise-linear limiter function

$$\mathcal{I}_\Delta(x) = \begin{cases} x & |x| \leq \Delta \\ \Delta\ \mathrm{sgn}(x) & \text{otherwise} \end{cases}\ . \tag{23}$$

Although the ML estimate (22) is asymptotically efficient, *i.e.*, its MSE converges to the Cramér-Rao bound (9) for large enough N, convergence in not uniform in A [3].

For any system with $M > 2$ in this Gaussian scenario, the ML estimate (21) can be obtained via an EM algorithm. In particular, by using the unquantized measurements $x[n] = A + \alpha[n]$ as the complete data we obtain [3]

$$\hat{A}_{\mathrm{EM}}^{(k+1)} = \mathcal{I}_\Delta\left(\hat{A}_{\mathrm{EM}}^{(k)} + \sum_{m=1}^{M} \frac{\sigma_\alpha\,\mathcal{K}_{Y_m}\left(\mathbf{y}^N\right)}{\sqrt{2\pi}\,N}\ \frac{\exp\left(-\frac{\left(X_{m-1}-\hat{A}_{\mathrm{EM}}^{(k)}\right)^2}{2\sigma_\alpha^2}\right) - \exp\left(-\frac{\left(X_m-\hat{A}_{\mathrm{EM}}^{(k)}\right)^2}{2\sigma_\alpha^2}\right)}{Q\left(\frac{X_{m-1}-\hat{A}_{\mathrm{EM}}^{(k)}}{\sigma_\alpha}\right) - Q\left(\frac{X_m-\hat{A}_{\mathrm{EM}}^{(k)}}{\sigma_\alpha}\right)}\right), \tag{24}$$

initialized with $\hat{A}_{\mathrm{EM}}^{(0)} = 0$. Provided that the log-likelihood function does not possess multiple local minima, (24) provides the ML estimate (21), *i.e.*, $\hat{A}_{\mathrm{ML}}\left(\mathbf{y}^N;\Delta\right) = \lim_{k\to\infty}\hat{A}_{\mathrm{EM}}^{(k)}$. Empirical evidence suggests that $\lim_{k\to\infty}\hat{A}_{\mathrm{EM}}^{(k)}$ is also asymptotically efficient.

4.2 Known Control Inputs

The ML estimate of A from $\mathbf{y}^N$ given a control vector $\mathbf{w}^N$, where $w[n]$ is a K-periodic sequence and N is a multiple of K, is provided by the following EM algorithm [3],

$$\hat{A}_{\text{EM}}^{(k+1)} = \mathcal{I}_\Delta \left(\hat{A}_{\text{EM}}^{(k)} + \sum_{\substack{1 \le \ell \le K \\ 1 \le m \le M}} \frac{\sigma_v \mathcal{K}_{Y_m}\left(\mathbf{y}^{\bar{N}}[\ell]\right)}{\sqrt{2\pi} N} \frac{\exp\left(-\frac{\left(X_{m-1} - \hat{A}_{\text{EM}}^{(k)} - w[\ell]\right)^2}{2\sigma_v^2}\right) - \exp\left(-\frac{\left(X_m - \hat{A}_{\text{EM}}^{(k)} - w[\ell]\right)^2}{2\sigma_v^2}\right)}{Q\left(\frac{X_{m-1} - \hat{A}_{\text{EM}}^{(k)} - w[\ell]}{\sigma_v}\right) - Q\left(\frac{X_m - \hat{A}_{\text{EM}}^{(k)} - w[\ell]}{\sigma_v}\right)} \right)$$

and $\hat{A}_{\text{ML}} = \lim_{k\to\infty} \hat{A}_{\text{EM}}^{(k)}$, initialized with $\hat{A}_{\text{EM}}^{(k)} = 0$, and where $\bar{N} = N/K$ and $\mathbf{y}^{\bar{N}}[\ell]$ is the $\bar{N} \times 1$ vector comprised of the elements of the ℓ-th K-decimated subsequence, *i.e.*,

$$\mathbf{y}^{\bar{N}}[\ell] \triangleq \left[y[\ell] \quad y[K+\ell] \quad \cdots \quad y[N-K+\ell] \right]^T \qquad \ell = 1, 2, \cdots, K \,. \tag{25}$$

Empirical evidence suggests that this EM estimate is also asymptotically efficient.

4.3 Control Inputs in the Presence of Feedback

In Sec. 3.3 we developed bounds on the worst-case information loss, by using control inputs that depend on A. Although these control inputs are not permissible, they suggest viable control input selection methods based on past quantized observations. For instance, for the system with $M = 2$, if $\hat{A}[n]$ is any consistent estimator of A based on $\mathbf{y}^n$, a reasonable choice for the control input sequence is as follows

$$w[n] = -\hat{A}[n-1] \,. \tag{26}$$

If the control sequence is selected according to (26), the ML estimate $\hat{A}_{\text{ML}}[n]$ of A based on $\mathbf{y}^n$ for $n = 1, 2, \cdots$ can be obtained using the following EM algorithm [3],

$$\hat{A}_{\text{EM}}^{(k+1)}[n] = \mathcal{I}_\Delta \left(\hat{A}_{\text{EM}}^{(k)}[n] + \sum_{m=1}^{n} \frac{\sigma_v \, y[m]}{\sqrt{2\pi}\, n} \frac{\exp\left(-\frac{\left(\hat{A}_{\text{ML}}[m-1] - \hat{A}_{\text{EM}}^{(k)}[n]\right)^2}{2\sigma_v^2}\right)}{Q\left(y[m] \, \frac{\hat{A}_{\text{ML}}[m-1] - \hat{A}_{\text{EM}}^{(k)}[n]}{\sigma_v}\right)} \right), \tag{27}$$

initialized with $\hat{A}_{\text{EM}}^{(0)}[n] = \hat{A}_{\text{ML}}[n-1]$ and $\hat{A}_{\text{ML}}[0] = 0$, where $\hat{A}_{\text{ML}}[n] = \lim_{k\to\infty} \hat{A}_{\text{EM}}^{(k)}[n]$. Empirical evidence suggests that (27) achieves the 2 dB bound (20) for any A in $(-\Delta, \Delta)$ for moderate N. Algorithms that achieve the bound (20), but require significantly fewer computations than (27), are developed in [3], as are extensions for systems with $M > 2$.

5 Multiple Sensors

In this section we consider the following network generalization of the estimation problem (1)-(2), where an unknown parameter A is to be estimated from observation of

$$y_i[n] = F(A + v_i[n] + w_i[n]) \qquad n = 1, 2, \cdots, N, \quad i = 1, 2, \cdots, L \tag{28}$$

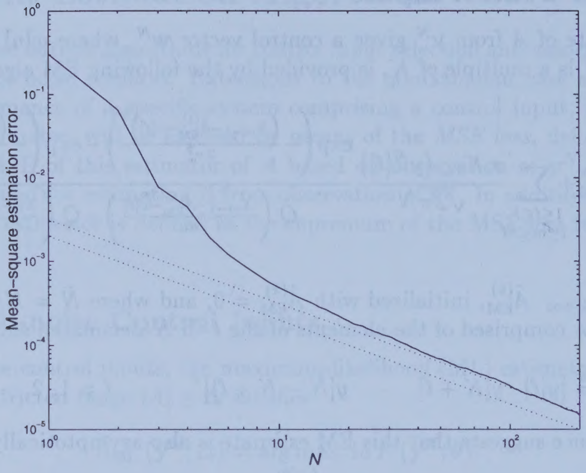


Figure 4: MSE for $\hat{A}_{\mathrm{ML}}[N]$ (solid curve) for a network of $L = 5$ two-level quantizers, exploiting feedback in the control input selection, and associated Cramér-Rao bounds (lower dotted line: bound for estimating A via s^N; upper dotted line: 2 dB bound). The sensor noise levels are 0.08, 0.08 0.08, 0.2, and 0.4, while $A = 0.4$ and $\Delta = 1$.

where $F(\cdot)$ is given by (2), the $v_i[n]$'s are independent IID Gaussian processes, and the $w_i[n]$'s denote the applied control input sequences. Such networks provide attractive models for a number of distributed estimation systems for which the possibly remote sensors are not co-located.

Straightforward extensions of the single-sensor systems developed in Secs. 3 and 4 yield appealing network generalizations that can be analyzed by means of the tools developed for the single-sensor case. For instance, we may consider networks exploiting feedback from the quantized outputs in the control input selection. The control input of each sensor at time n can be selected using (26), where $\hat{A}[n-1]$ denotes the estimate based on observations collected from all L sensors up to and including time $n-1$. In Fig. 4 we show the MSE performance of the network extension of the ML estimator (27) for $L = 5$. As in the single-sensor case, the MSE of this estimator practically achieves the bound corresponding to a 2 dB information loss for moderate N values.

References

[1] H. C. Papadopoulos, *Analysis and Applications of a Class of Signal Quantizers.* PhD thesis, M.I.T., Cambridge, MA. In preparation.

[2] R. Benzi, A. Sutera, and A. Vulpiani, "The mechanism of stochastic resonance," *J. Phys.*, vol. A14, pp. L453–L457, 1981.

[3] H. C. Papadopoulos, G. W. Wornell, and A. V. Oppenheim, "Signal encoding from noisy measurements using quantizers with dynamic bias control," preprint.

Decision Feedback Sequence Estimation for Unwhitened ISI Channels

Abdulrauf Hafeez[1] and Wayne E. Stark
Dept. of Elect. Engineering and Computer Science
1301 Beal Ave. , University of Michigan, Ann Arbor, MI 48109-2122.
E-mail: rauf@eecs.umich.edu, stark@eecs.umich.edu.

Abstract– In this paper, a form of decision-feedback sequence estimation (DFSE) is proposed for unwhitened inter-symbol interference (ISI) channels. DFSE which is a reduced-complexity alternative to maximum likelihood sequence estimation, performs very well on whitened (minimum-phase) ISI channels. However, it may not be feasible to compute the whitening filter for many time-varying channels like the multiuser asynchronous direct-sequence CDMA channel. On such channels, it is desirable to use the Ungerboeck formulation which operates directly on discrete-time unwhitened statistic obtained from conventional matched filtering. Unfortunately, decision-feedback sequence estimation for unwhitened channels is interference limited because the decision statistic depends on both past and future inputs. Specifically, the decision rule that chooses the path with the best accumulated metric leading to a reduced state, fails to account for interference from the future inputs not included in the reduced state. We propose a modified decision rule to be integrated into the Ungerboeck form of sequence estimation which is not sensitive to channel phase. It can be implemented as a single stage or a multistage scheme depending on the amount of dispersion present in the channel. Simulation results indicate that the algorithm used on unwhitened channels with low to moderate dispersion, provides similar performance/complexity tradeoffs as the DFSE used on the corresponding whitened minimum-phase channels. The algorithm is especially attractive for multiuser asynchronous DS-CDMA channels with long spreading codes.

1 Introduction

There are two main approaches to maximum likelihood sequence estimation (MLSE) for known inter-symbol interference (ISI) channels with additive noise. Forney's approach [1] consists of using the Viterbi algorithm on the sampled output of whitened matched filters while Ungerboeck's formulation [2] operates directly on conventional matched filter output samples without the need of noise whitening. Due to the exponential complexity of these methods, several low-complexity sub-optimal schemes have been developed. A promising reduced-complexity alternative to MLSE using Forney's approach is decision feedback sequence estimation (DFSE) [3], [4]. The scheme provides an excellent tradeoff between performance and complexity.

---

[1] This work was supported in part by the Army Research Office under grant DAAH04-95-I-0246 and by Ericsson Inc.

ISI may be the result of spectral shaping at the transmitter and/or distortion in the physical channel. Moreover, multiuser interference can also be viewed as a form of time-varying ISI. A receive filter matched to a non-Nyquist transmit pulse-shaping filter or to the overall channel impulse response, gives rise to bi-directional ISI and correlated noise. The noise has to be whitened in order to use DFSE. In case of rapidly (or periodically) time-varying ISI channels, it may not be feasible to track the whitening filter (the computation involves polynomial factorization and inversion). An example is the multiuser asynchronous direct-sequence CDMA channel where the spreading of users' transmit pulses with long non-orthogonal spreading codes gives rise to rapidly time-varying ISI over the composite data sequence. Clearly, Ungerboeck's form is better suited for such channels as noise whitening is expensive [5]. However, a reduced-state Ungerboeck variant (which we call UDFSE[2]) similar to the DFSE does not work as well. The problem arises from the nature of the unwhitened statistic which depends on both past and future input symbols. Consequently the decision rule borrowed from DFSE, which decides in favor of the path with the best accumulated (Euclidean distance) metric, leading to a reduced state, ignores interference from the future inputs not included in the reduced state. As a result, the decisions obtained in UDFSE are interference-limited even in the absence of feedback errors. The M-algorithm [7], used on unwhitened statistic, suffers from the same problem as paths are selected based on accumulated metrics which do not reflect all relevant interference components.

In the case of UDFSE, where the reduced state is pre-determined, the interfering anti-causal components are known. Thus the decision rule can be modified to nullify the effect of these components recursively. This can be accomplished in two ways: by using conventional hard decisions on the relevant anti-causal interference components or by using decisions obtained at the output of the preceding stage in a multistage scheme. In either case, the preliminary decisions do not enter the branch metrics directly. The number of stages to use depends on the amount of dispersion present in the channel. Since conventional hard decisions are used to cancel interference partially, the modified scheme may not be suitable for highly dispersive channels.

Simulation is undertaken to compare the performance of various algorithms on single-user ISI channels and multiuser asynchronous CDMA channels. It is found that the modified UDFSE scheme obtains sizable gains over UDFSE on channels with low to moderate amount of dispersion. On such channels, it provides similar performance/complexity tradeoff as DFSE used on the corresponding whitened minimum-phase channels. The proposed scheme performs very well on DS-CDMA channels as the correlations among user's spreading codes are small. It is noted that a single-stage algorithm suffices when the dispersion is low while a second stage is needed in medium dispersion. On highly dispersive channels, the performance is not that good and may be worse than UDFSE.

The paper is organized as follows. We present the system model in section 2. The two approaches to MLSE are reviewed in section 3. In section 4, we describe various methods of decision-feedback sequence estimation in detail and develop the modified UDFSE algorithm. Simulation results are presented in section 5 where the performance of the various schemes is compared for single user ISI channels and multiuser DS-CDMA channels.

---

[2] The scheme is referred to as Delayed Decision Feedback Sequence Estimation with *a standard matched filter* in [6]

2 System model

Consider the transmission of linearly-modulated digital data over a (time-varying) time-dispersive additive white Gaussian noise channel. Assume that the receiver has perfect knowledge of the carrier phase, symbol timing and the impulse response of the channel (assumed constant for symbol period T). After coherent down-conversion, the receiver employs a filter $h_n(t)$ matched to the cascade of the transmit pulse-shaping filter and the channel impulse response and samples the output at symbol spaced intervals. The sequence of sampled matched filter outputs is a sufficient statistic for estimating the transmitted data sequence. Thus an equivalent discrete-time channel model is obtained.

The matched filter output at time n is given by

$$y_n = \sum_{i=-L}^{L} s_{n,i} b_{n-i} + z_n \tag{1}$$

where $\{b_n\}$ is the transmitted data sequence assumed to be i.i.d. $b_n \in A_M$ (an M-ary alphabet). $s_{n,i}$ are channel correlations sampled at time n, given by

$$s_{n,i} = \begin{cases} h_n(t) * h_n(-t)|_{t=iT} & |i| \leq L \\ 0 & |i| > L \end{cases} \tag{2}$$

where $L < \infty$. $\{z_n\}$ is a complex Gaussian noise process with mean 0 and covariance $\mathrm{E}[z_n^* z_{n-i}] = N_0 s_{n,i}$. We will refer to the above model as the unwhitened model.

Let $s_n(D) = \sum_i s_{n,i} D^i$, then an equivalent discrete-time white Gaussian noise model is obtained by noting

$$s_n(D) = f_n^*(D^{-1}) f_n(D)$$

where $f_n(D) = \sum_{i=0}^{L} f_{n,i} D^i$ has all its roots inside the unit circle (minimum-phase channel). Thus the statistic obtained when $\{y_n\}$ is filtered by the anti-causal whitening filter $1/f_n^*(D^{-1})$, is given by

$$y'_n = \sum_{i=0}^{L} f_{n,i} b_{n-i} + z'_n \tag{3}$$

$\{z'_n\}$ is a proper[3] complex white Gaussian noise process with mean 0 and variance N_0.

3 Maximum likelihood sequence estimation

The maximum likelihood sequence estimator (MLSE) determines as the most likely sequence transmitted, the sequence b^N that maximizes the metric [2]

$$\mathrm{U}\left(y^N, b^N\right) = \sum_{n=1}^{N} 2\mathrm{Re}\{b_n^* y_n\} - \sum_{i=1}^{N}\sum_{k=1}^{N} b_i^* s_{i,k} b_k. \tag{4}$$

where $x^n = x_1, x_2, \ldots, x_n$. Under the condition:

$$s_{n,i} = s_{n-i,-i}^* \text{ or equivalently } s_{n,-i} = s_{n+i,i}^*, \tag{5}$$

it can be shown that the above metric can be computed recursively as

$$\mathrm{U}(y^n, b^n) = \mathrm{U}(y^{n-1}, b^{n-1}) + \mathrm{V}(y_n, b_n, \sigma_{n-1}) \tag{6}$$

---

[3] $\mathrm{E}[\mathrm{Re}\{z'_n\}\mathrm{Im}\{z'_n\}] = 0$

where V(.) is known as the branch metric, given by

$$\mathrm{V}(y_n, b_n, \sigma_{n-1}) = \mathrm{Re}\left\{ b_n^* \left(2y_n - s_{n,0} b_n - 2\sum_{i=1}^{L} s_{n,i} b_{n-i} \right) \right\} \tag{7}$$

and $\sigma_n = b_n, b_{n-1}, \ldots, b_{n-L+1}$ is the state at time n.

The above is a more generalized form of the Ungerboeck formulation (where $s_{n,i} = s_i$ i.e. the channel is fixed) independently derived in [8] in detail. The condition in (5) holds for time-varying channels like the asynchronous DS-CDMA channel. The application of this algorithm to the multiuser sequence estimation problem was noted by Verdu [9].

The maximum likelihood sequence can also be determined equivalently by applying the Viterbi algorithm on the whitened statistic $\{y'_n\}$ as shown by Forney [1]. The metric in this case can be written as

$$\mathrm{U}'\left(y'^{N}, b^{N}\right) = \sum_{n=1}^{N} \left[2\mathrm{Re}\left(y'_n \sum_{i=0}^{L} f_{n,i}^* b_{n-i}^* \right) - \left\| \sum_{i=0}^{L} f_{n,i} b_{n-i} \right\|^2 \right] \tag{8}$$

which can be computed recursively as

$$\mathrm{U}'\left(y'^{n}, b^{n}\right) = \mathrm{U}'(y'^{n-1}, b^{n-1}) + \mathrm{V}'(y'_n, b_n, \sigma_{n-1}) \tag{9}$$

where

$$\mathrm{V}'(y'_n, b_n, \sigma_{n-1}) = 2\mathrm{Re}\left(y'_n \sum_{i=0}^{L} f_{n,i}^* b_{n-i}^* \right) - \left\| \sum_{i=0}^{L} f_{n,i} b_{n-i} \right\|^2 . \tag{10}$$

The two algorithms have the same order of computational complexity and storage requirement proportional to M^L.

4 Decision feedback sequence estimation

A reduced-complexity sub-optimal alternative to MLSE using the Viterbi algorithm is decision feedback sequence estimation DFSE [3], [4]. A parameter called memory order J is chosen arbitrarily smaller than the channel memory L and the trellis is collapsed into M^J states corresponding to the J most recent hypothesized symbols only. Since the reduced state falls short in providing all the information needed to compute branch metrics for the next iteration, the algorithm uses the path with the best accumulated metric, leading to each state, to extract the rest of the information required. In other words, the algorithm makes tentative conditional decisions on the input symbols more than J samples in the past (given the J most recent hypothesized input symbols) to cancel their effect on the current output sample. Note that this algorithm operates on whitened statistic $\{y'_n\}$ which depends on past and present inputs only.

4.1 UDFSE

In this section, we consider decision-feedback sequence estimation in the Ungerboeck form, referred to here as UDFSE (derived in [6]).

Let $\beta_n = b_n, b_{n-1}, \ldots, b_{n-J+1}$ represent the reduced state at time n and $\mathrm{m}(\beta_n)$ be the associated path metrics, defined as

$$\mathrm{m}(\beta_n) \triangleq \max_{b^{n-J}} \mathrm{U}(y^n, b^n) \tag{11}$$

then the algorithm can be summarized as:

$$\mathrm{m}(\beta_n) = \max_{b_{n-J}}[\mathrm{m}(\beta_{n-1}) + \mathrm{W}(y_n, b_n, \beta_{n-1})] \tag{12}$$

$$\mathrm{W}(y_n, b_n, \beta_{n-1}) = \mathrm{Re}\left\{ b_n^* \left(2y_n - s_{n,0}b_n - 2\sum_{i=1}^{J} s_{n,i}b_{n-i} - 2\sum_{i=J+1}^{L} s_{n,i}\hat{b}_{n-i}(\beta_{n-1}) \right) \right\} \tag{13}$$

$\{\hat{b}(\beta_n)\}$ are tentative conditional decisions on inputs more than J samples in the past, obtained from the path with the best accumulated metric, leading to reduced state β_n.

$$\hat{b}_{n-J}(\beta_n) = \arg\max_{b_{n-J}}[\mathrm{m}(\beta_{n-1}) + \mathrm{W}(y_n, b_n, \beta_{n-1})] \tag{14}$$

A drawback of UDFSE: Although both full-state formulations in section 3 yield the maximum likelihood sequence estimate, the corresponding reduced-state derivations perform quite dissimilarly. In fact, UDFSE does not perform nearly as well as DFSE on most channels. The problem arises from the nature of unwhitened statistic. Recall that unlike whitened statistic, unwhitened statistic $\{y_n\}$ depends on both past and future input symbols. Now note from (14) that when deciding (conditionally) b_{n-J} in the n-th step, the UDFSE algorithm has ignored interference from inputs $b_{n+1}, b_{n+2}, \ldots, b_{n-J+L}$ which directly affect output sample y_{n-J} corresponding to b_{n-J}. This means that the decision rule of (14), used to decide b_{n-J}, based on the knowledge of only J inputs in the future is inherently near-sighted unless it has some sort of side information available about another $L-J$ inputs in the future. Thus unlike the whitened channel counterpart, UDFSE suffers from a major drawback: the conditional decisions are interference-limited from future inputs even in the absence of feedback errors. The M-algorithm which operates on unwhitened statistic is also plagued by the same problem.

4.2 Modified UDFSE

When the reduced-state is pre-determined as in the case of UDFSE, the problem can be alleviated by modifying the decision rule to include a bias for cancelling the effect of the residual interference. Note that the full-state algorithm does not decide b_{n-J} until the $(n-J+L)$th step when the interference from the relevant future inputs has been averaged out. On the other hand, the reduced-state scheme accounts for the contribution of only J inputs in the future. The bias needed to cancel the residual anti-causal interference completely in the reduced-state scheme can be found by comparing UDFSE term by term against the full-state algorithm.

The modified algorithm (MUDFSE) can be outlined as follows: Path metrics are calculated as in UDFSE using equation(12). Conditional decisions are made (and the corresponding survivor paths are chosen) using the modified decision rule:

$$\hat{b}_{n-J}(\beta_n) = \arg\max_{b_{n-J}}[\mathrm{m}(\beta_{n-1}) + \mathrm{W}(y_n, b_n, \beta_{n-1}) + \mathrm{bias}(b_{n-J})] \tag{15}$$

where

$$\mathrm{bias}(b_n) = -2\mathrm{Re}\left\{ b_n^* \sum_{i=1}^{L-J} s_{n,-J-i}\tilde{b}_{n+J+i} \right\}, \quad \tilde{b}_n = \mathrm{sign}(y_n). \tag{16}$$

Note that tentative hard decisions $\{\tilde{b}_n\}$ (which may not be very reliable) affect branch metrics (12) only through conditional decisions $\{\hat{b}(\beta_n)\}$. Also note that the bias term does not add significantly to the computational load and storage requirement of the algorithm which is on the order of $(G-J)M^J$ (same as DFSE), where G is the decision lag. The algorithm is delayed by $L-J$ samples as y_{n+L-J} is needed in the nth step.

4.3 Multistage MUDFSE

MUDFSE can be run in a multistage configuration where decisions obtained at the output of the first stage are fed back to compute the bias in the second stage and so on. i.e.

$$(\tilde{b}_n)_1 = \text{sign}(y_n) \text{ and } (\tilde{b}_n)_i = (\bar{b}_n)_{i-1}, \ i > 1$$

where $\{(\bar{b}_n)_i\}$ are decisions obtained from the ith stage at lag G_i ($G_i >> L$). Note that decisions $\{(\bar{b}_n)_1\}$ are likely to be much more reliable than $\{(\tilde{b}_n)_1\}$. The complexity and delay of a Q-stage scheme is given by $\sum_{i=1}^{Q}(G_i - J_i)M^{J_i}$ and $\sum_{i=1}^{Q} G_i + L - J_1$ respectively.

Since the reliability of symbol-by-symbol decisions $\{\text{sign}(y_n)\}$ depends heavily on the amount of dispersion present in the channel, the use of these decisions in MUDFSE is warranted for channels with low-to-medium dispersion only. On highly dispersive channels, it may be better to ignore the anti-causal interference rather than try to cancel it with unreliable decisions. On such channels, UDFSE may be the preferable scheme (or the first stage followed by MUDFSE as the second stage, in a two-stage setup).

5 Performance results

Performance is evaluated via simulation. We consider binary signaling over two static time-dispersive additive white Gaussian noise channels. Channel 1 is given by $f = (0.8614394, 0.2584344, -0.1000742, -0.2740468, 0.130008, 0.1000742, -0.0375511, 0.1122497, -0.1137419, -0.22837)$ and channel 2 is given by $f = (0.7071067, -0.153016, 0.4679421, -0.306526, 0.2846515, -0.1041423, 0.2437057, 0.0679643, -0.0705263, 0.052918)$. Both channels are minimum phase and are arbitrarily chosen. The receiver has perfect knowledge of the symbol timing and the impulse response of the channel.

Figures 1 and 2 show the bit-error rate performance of various schemes on channels 1 and 2 respectively. Each simulation was run for a count of 1000 errors. The various schemes are indexed with the memory order and the decision lag (J, G). Except DFSE, all algorithms operate on symbol-spaced samples obtained at the output of conventional matched filters. Ideal noise whitening is assumed for DFSE. Channel 1 has low-to-moderate amount of dispersion. All schemes shown in Fig 1, except MLSE, have the same order of overall complexity and storage requirement. Fig 1 shows that $(5, 45)$ Modified UDFSE gains $0.5 - 1.5$ dB over $(5, 45)$ UDFSE. Another $0.5 - 2.25$ dB is gained by the two-stage scheme $[(4, 45), (4, 45)]$ 2MUDFSE over single-stage $(5, 45)$ MUDFSE. No significant amount of further gain is achieved by the three-stage scheme with the same overall complexity. Its performance, however, is almost identical to that of $(5, 45)$ DFSE on this channel. Note that, here we have neglected the effects of non-ideal noise-whitening on the performance of DFSE, resulting from fixed delay constraint. The additional delay required can be compared to the delay of a multi-stage scheme. Moreover, the performance of DFSE depends heavily on channel phase with the best performance achieved on minimum-phase channels. On the contrary, the performance of receiver structures that operate on unwhitened statistic is independent of channel phase.

Channel 2 is more dispersive than channel 1 (the main signal element f_0 contains only 50% of the total energy). Fig 2 shows that the modified single-stage scheme $(5, 45)$ MUDFSE is actually a bit worse than $(5, 45)$ UDFSE for most of the SNR range. This is because the conventional hard decisions used in the modified scheme to compute the bias term, are somewhat unreliable due to heavy dispersion. However, the two-stage scheme $[(4, 45), (4, 45)]$ 2MUDFSE, with the same overall complexity as the single-stage algorithms. manages to gain about half a dB owing to the much improved reliability of decisions available to the second stage. Its performance is still far from that achieved by $(5, 45)$ DFSE especially at low noise. The two-stage scheme with higher memory order, however, compares more favorably with $(5, 45)$ DFSE.

Next, we simulate a BPSK modulated asynchronous DS-CDMA system with eight users whose signature waveforms are derived from Gold sequences of length 31. The relative delays of users are fixed for the simulation and are in an increasing order. The channel spectrum $S(D) = S_1^T D^{-1} + S_0 + S_1 D$ is given by

$$\frac{1}{31}\begin{bmatrix} 31 & 8-D & 7 & 6+D & -3+2D & -6+5D & -3+2D & -2+D \\ 8-D^{-1} & 31 & -D & 9-2D & -6-3D & -7-2D & -D & 9-2D \\ 7 & -D^{-1} & 31 & -D & -1 & -D & -1 & -2+D \\ 6+D^{-1} & 9-2D^{-1} & -D^{-1} & 31 & -D & -1 & -D & -1 \\ -3+2D^{-1} & -6-3D^{-1} & -1 & -D^{-1} & 31 & -D & -1 & -D \\ -6+5D^{-1} & -7-2D^{-1} & -D^{-1} & -1 & -D^{-1} & 31 & -D & -1 \\ -3+2D^{-1} & -D^{-1} & -1 & -D^{-1} & -1 & -D^{-1} & 31 & -D \\ -2+D^{-1} & 9-2D^{-1} & -2+D^{-1} & -1 & -D^{-1} & -1 & -D^{-1} & 31 \end{bmatrix}$$

The equivalent periodically time-varying ISI channel can be obtained from the relation

$$s_{\kappa(n),i} = \begin{cases} S_1(\kappa(n), \kappa(n) - i + K) & \kappa(n) - i < 1 \\ S_0(\kappa(n), \kappa(n) - i) & 1 \leq \kappa(n) - i \leq K \\ S_1(\kappa(n) - i - K, \kappa(n)) & \kappa(n) - i > K \end{cases} \tag{17}$$

for $|i| \leq K - 1$, where K is the number of users and $\kappa(n)$ is the user index given by $\kappa(n) = n \bmod K$.

Figure 3 shows the bit-error rate of user 1 for various schemes when all users have identical signal-to-noise ratio. Each simulation was run for a count of 500 errors. It is evident that even with ideal power control, the performance of the conventional single-user detector is in no comparison with optimum MLSE. With four states, $(2, 28)$ UDFSE and (4) M-algorithm that operate directly on the matched filter output, do not improve much over the conventional detector. Linear-decorrelator that nulls out all interference, loses about $0.7 - 1.0$ dB compared to MLSE. $(2, 28)$ DFSE and (4) M-algorithm that operate on the equivalent whitened minimum-phase channel, closely approach MLSE performance. The proposed single-stage and two-stage modified UDFSE algorithms give by far the best performance on this channel (next to MLSE). This behavior is attributable to low channel cross-correlations typical of well-designed spreading codes (compare to the channel in figure 1). The second stage of 2MUDFSE does not provide any significant gain over the decisions obtained in the first stage. $(2, 28)$ MUDFSE (with four states) is only 0.3 dB worse than MLSE (128 states) and gains $1.0 - 3.0$ dB over UDFSE.

Figure 4 shows the bit-error rate of user 1 against the signal-to-noise ratio of the rest of the users. SNR of user 1 is held constant at 10.0 dB. It can be seen that the conventional detector, UDFSE and the unwhitened M-algorithm are limited by multiuser interference. Linear-decorrelator is ideally near-far resistant which means that its performance is the same as that of MLSE under worst case interference. In other conditions, DFSE and the whitened M-algorithm perform better. The proposed modified UDFSE schemes outperform the other methods and converge to MLSE in high multiuser interference.

Note that MUDFSE is especially suitable for long spreading codes used in today's cellular systems. Channel factorization and inversion operations are difficult to perform on the time-varying channels that result from long spreading codes and arrival and departure of users. However, partial cross-correlations of long codes can be easily generated locally at the receiver (especially at the uplink).

6 Conclusions and future work

We described a form of decision-feedback to be integrated into the Ungerboeck formulation for sequence estimation on ISI channels. The resulting reduced-complexity algorithm provides similar performance/complexity tradeoffs as the decision feedback sequence estimator on low to moderately dispersive channels. It, however, is insensitive to channel phase and does not require noise-whitening which makes it attractive for time-varying channels like the asynchronous DS-CDMA channel.

The proposed scheme requires knowledge of received signal amplitudes. Thus, it is imperative that an adaptive form of the algorithm be developed to use on fading channels. In [10], UDFSE was considered for equalization in the global system for mobile communications (GSM) receiver. It was found that UDFSE with four states provides adequate performance with some degradation for rural area, typical urban and hilly terrain channels with six rays but exhibits inferior performance for a uniform ray power profile. It would be interesting to consider the proposed algorithm for these channels.

Acknowledgment: The authors will like to acknowledge CACC, North Carolina State University for the use of their facilities for this research. Moreover, the authors will like to thank Prof. A. Duel-Hallen and Mr. A. El-Ezaby of NCSU and Dr. G. E. Bottomley of Ericsson Inc. for useful comments and suggestions.

References

[1] G. D. Forney, "Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference", *IEEE Trans. Information Theory*, Vol. IT-18, No. 3, pp. 363-378, May 1972.

[2] G. Ungerboeck, "Adaptive Maximum-Likelihood Receiver for Carrier-Modulated Data-Transmission Systems", *IEEE Trans. Comm. Technology*, Vol. COM-22, No. 5, pp. 624-636, May 1974.

[3] A. Duel-Hallen and C. Heegard, "Delayed Decision-Feedback Sequence Estimation", *IEEE Trans. Comm.*, Vol. 37, No. 5, pp. 428-436, May 1989.

[4] V. M. Eyuboglu and S. U. Qureshi, "Reduced-state Sequence Estimation with Set Partitioning and Decision Feedback", *IEEE Trans. Comm.*, pp. 13-20, Jan. 1988.

[5] L. Wei and L. K. Rasmussen, "A Near Ideal Noise Whitening Filter for an Asynchronous Time-Varying CDMA System", *IEEE Trans. Comm.*, Vol. 44, No. 10, pp. 1355-1361, Oct. 1996.

[6] A. Duel and C. Heegard, "Delayed Decision Feedback Sequence Estimation for QAM and Trellis Coded Systems", *20th Conf. Info. Sc. and Sys.*, pp. 367-372, Mar. 1986.

[7] J. B. Anderson and S. Mohan, "Sequential Coding Algorithms: A Survey Cost Analysis", *IEEE Trans. Comm.*, Vol. COM-32, pp. 169-176, Feb. 1984.

[8] G. E. Bottomley and S. Chennakeshu, "Adaptive MLSE Equalization Forms for Wireless Communications", *Fifth Virginia Tech. Symp. Wireless Personal Comm.*, pp. 20-1-20-12, May 1995.

[9] S. Verdu, "Minimum Probability of Error for Asynchronous Gaussian Multiple-Access Channels", *IEEE Trans. Info. Th.*, Vol. IT-32, No. 1, pp. 85-96, Jan. 1986.

[10] K. Wesolowski, R. Krenz and K. Das, "Efficient Receiver Structure for GSM Mobile Radio", *Int. J. Wireless Information Networks*, Vol. 3, No. 2, 1996.

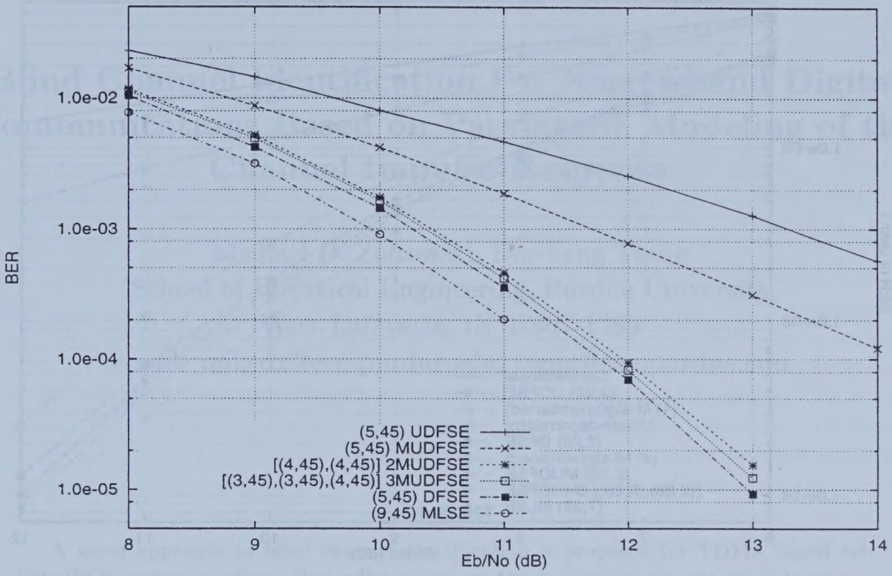


Figure 1: BER performance of various detection schemes on channel 1.

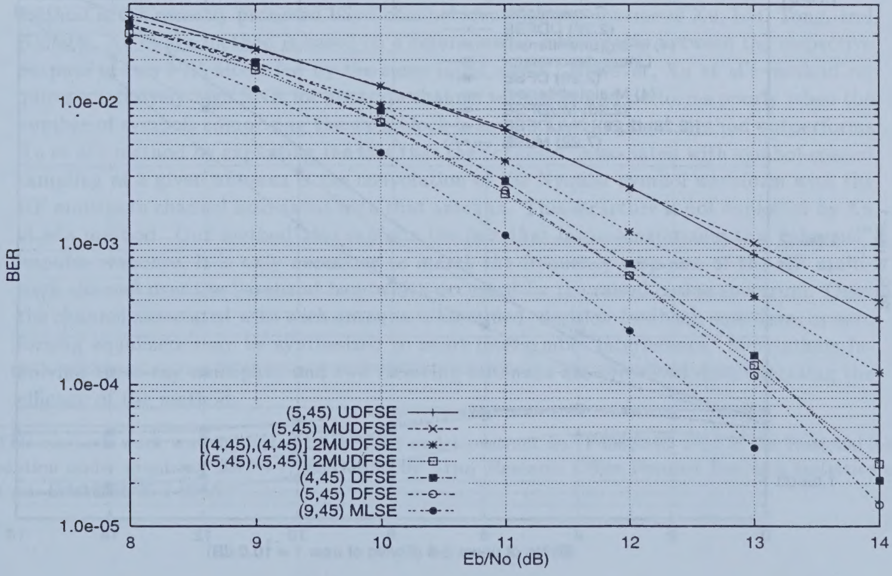


Figure 2: BER performance of various detection schemes on channel 2.

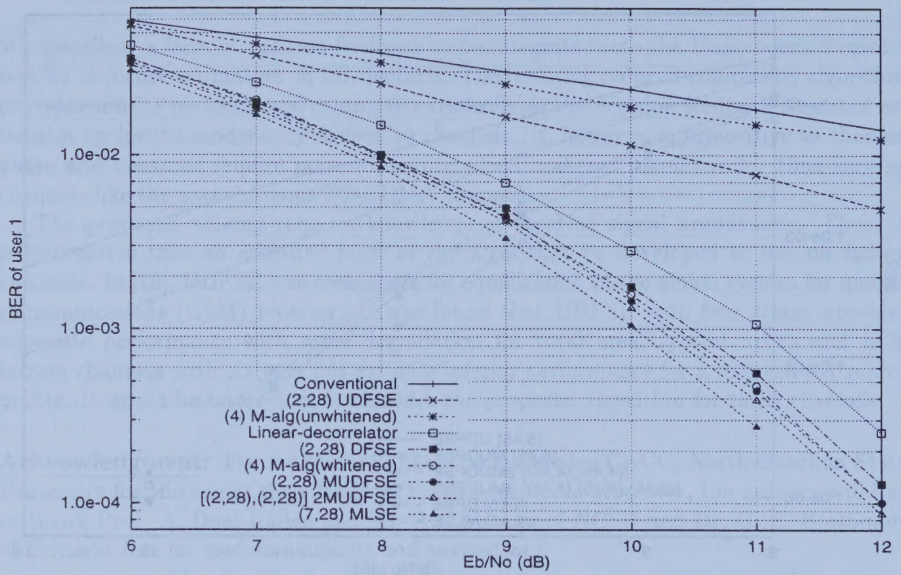


Figure 3: BER performance of various detection schemes on an 8-user asynchronous DS-CDMA channel.

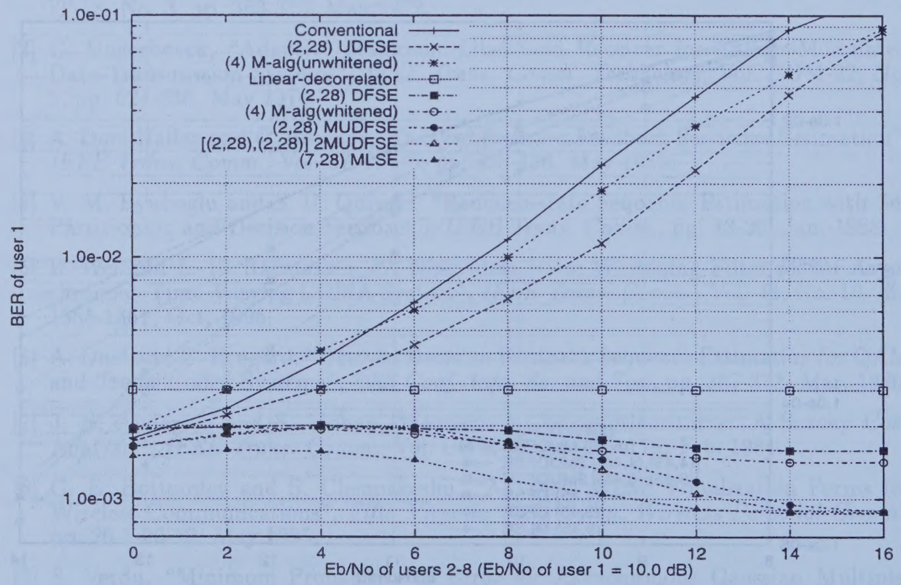


Figure 4: Near-far performance of various detection schemes on an 8-user asynchronous DS-CDMA channel.

Blind Channel Identification For Narrowband Digital Communications Based on Parametric Modeling of the Channel Impulse Response [†]

Michael D. Zoltowski, Der-Feng Tseng
School of Electrical Engineering, Purdue University
West Lafayette, IN 47907-1285
email: mikedz@ecn.purdue.edu, tseng@ecn.purdue.edu

Abstract

A novel approach to blind channel identification is proposed for TDMA based cellular/PCS communications that relies solely on the respective outputs of at least two spatially separated antennas. The proposed scheme holds the greatest promise in terms of the trade-off amongst convergence rate, computational complexity, robustness to model mismatch, and symbol error performance for TDMA systems where the multipath time delay spread is less than (or on the order of) the null-to-null mainlobe of the pulse symbol waveform, as is the case with the IS-136 standard. The starting point for the proposed method is the recently proposed blind channel identification scheme of Xu, Liu, Tong, and Kailath. Xu et al's method is based on a deterministic relationship between the respective outputs of two FIR filters fed by the same input signal. However, Xu et al's method requires a relatively high SNR for accurate channel estimates and performs poorly when the number of symbols comprising the data block is small. The proposed method outperforms Xu et al's method by exploiting the fact that the "channel" associated with symbol-spaced sampling at a given antenna is the convolution of the Nyquist symbol waveform with the RF multipath channel associated with that antenna. This structure is not exploited by Xu et al's method. Our method also exploits the fact that in parameterizing this "channel" impulse response, it is only necessary to match the frequency response of the RF multipath channel over the baseband bandwidth occupied by the raised cosine spectrum. Once the channel associated with each antenna is identified, decision feedback equalizers or zero forcing equalizers may be synthesized to avert intersymbol interference. Simulations involving three-ray multipath and two receiving antennas are presented demonstrating the efficacy of the method.

[†] This research work was supported by AFOSR under contract no. F49620-97-1-0275, the National Science Foundation under grant no. MIPS-9708309, and by Army Research Office Focused Research Initiative under grant no. DAAH04-95-1-0246.

1 Introduction

RF propagation in urban cellular environments involve multipath: reflections off buildings and other structures. Employing Nyquist pulses in narrowband digital communications, equalization is necessary when the multipath time delay spread is a significant fraction of the symbol waveform duration. The multipath delay spread in a typical urban environment has been measured to be 7-10 μs.

Recently there has been an intense research effort on blind channel identification and equalization. It is known that second-order statistical(SOS) approaches to blind channel identification [1], [2], [3] require less data samples compared with blind identification schemes based on higher-order statistics (HOS) [4], [5], [6]. However, the separation between the signal and noise subspaces is a sensitive task; small channel length mismatch can introduce large channel estimation errors. Moreover, SOS based blind identification schemes still perform rather poorly given only a relatively small number of symbols.

Xu et al[7] proposed an approach to identifying the channels that is not based on exploiting the statistical characteristics of the input. Xu et al's method is based on a deterministic relationship between respective outputs of two FIR filters fed by same input signal. Xu et al's method is the basis for the blind channel identification scheme proposed herein. We here show that the performance of Xu et al's method may be improved dramatically by exploiting the following a-priori information:

- the "channel" associated with symbol-spaced sampling at a given antenna is the convolution of the Nyquist symbol waveform with the RF multipath channel associated with that antenna
- the relative time delay of each multipath is the same at each antenna
- the "real" channel realized by symbol-spaced sampling starting at $t = 0$ **and** the "virtual" channel realized by symbol-spaced sampling starting at $t = \frac{T_o}{2}$ are structurally related
- the "channel" impulse response may be parameterized via a small number of "oversampled-spaced" taps encompassing the multipath delay spread in order to match the frequency response of the associated RF multipath channel over the baseband bandwidth occupied by the raised cosine spectrum
- the **same** sample-spaced coefficients characterize both the "real" and "virtual" channels

As the simulations presented in Section 4 show, exploiting this a-priori information leads to dramatically improved performance allowing Xu et al's method to accurately identify the channels under adverse multipath conditions with only 30 symbols.

2 Problem Statement

The transmitted signal consists of a burst of N_b modulated symbols, $s[m], m = 0, \ldots, N_b - 1$, multiplied by a symbol waveform, $p_{sr}(t)$, (for example, a pulse with a square-root raised cosine spectrum) with a symbol being transmitted every T_o seconds. The transmitted signal can be represented as:

$$m(t) = \sum_{m=0}^{N_b-1} s[m]p_{sr}(t - mT_o) \tag{1}$$

Assume the signal is received by an M-element antenna array. When the environment of the communication system is frequency selective, i.e., non-flat, the impulse response of the RF multipath channel associated with the reference antenna (assumed without loss of generality to be that antenna enumerated as antenna number one) is modeled as

$$h_{RF_1}(t) = \sum_{k=1}^{K} g_{1k}\delta_a(t - \tau_k). \tag{2}$$

where τ_k is the delay and g_{1k} is the gain and phase of the k-th multipath channel at antenna 1, and $\delta_a(t)$ is the Dirac delta function. The received signal excluding noise is $r_1(t) = m(t) * h_{RF_1}(t) = \sum_{k=1}^{K} g_{1k}m(t - \tau_k)$, where "$*$" denotes linear convolution. The received signal at antenna i, where $i = 1, 2, \ldots, M$ is

$$r_i(t) = \sum_{k=1}^{K} a_i(\theta_k)g_{1k}m(t - \tau_k) + w_i(t) \tag{3}$$

where $w_i(t)$ is assumed to be additive Gaussian noise with a flat power spectrum across the passband of the signal of interest. In vector form,

$$\mathbf{r}(t) = \sum_{k=1}^{K} \mathbf{a}(\theta_k)g_{1k}m(t - \tau_k) + \mathbf{w}(t) \tag{4}$$

where $\mathbf{a}(\theta_k)$ is the $M \times 1$ array manifold that describes the relative gains and phases induced across the array aperture due to a narrowband planewave arriving at angle θ (first component is unity.) Define $g_{ik} = a_i(\theta_k)g_{1k}$, then $r_i(t) = m(t) * h_{RFi}(t) + w_i(t)$ where

$$h_{RF_i}(t) = \sum_{k=1}^{K} g_{ik}\delta_a(t - \tau_k). \tag{5}$$

The received signal after front-end square-root spectral filtering, $y_i(t) = r_i(t)*p_{sr}(-t)$ $i = 1, 2, \cdots, M$, may be described as

$$\boxed{y_i(t) = \sum_{m=0}^{N_b-1} s[m]h_i(t - mT_o) + n_i(t)} \tag{6}$$

where

$$\begin{aligned} h_i(t) &= p_{sr}(t) * p_{sr}(-t) * h_{RF_i}(t) = p_{rc}(t) * h_{RF_i}(t) \\ &= \sum_{k=1}^{K} g_{ik}p_{rc}(t - \tau_k), \end{aligned} \tag{7}$$

and $p_{rc}(t) = p_{sr}(t) * p_{sr}(-t)$ has a raised cosine spectrum. The noise, $n_i(t) = w_i(t) * p_{sr}(t)$, has a raised cosine power spectral density.

In blind equalization, $s[m]$ and $h_i(t)$ are both unknown. We are here proposing a new approach to blind channel identification that is based on Xu et al's method[7], but takes advantage of other a-priori information to identify the channel with a smaller number of symbols.

3 Blind Channel Identification

Thomas and Zoltowski[9], [10] recently proposed a novel algorithm for determining the channels premised on matching the frequency responses over the bandwidth of the autocorrelation of the pulse symbol waveform. The approach proposed involves determining $h_{ci}(t)$ of duration roughly equal to the multipath time delay spread such that ideally

$$h_{ci}(t) * p_{rc}(t) = h_{RF_i}(t) * p_{rc}(t). \tag{8}$$

That is, if $p_{rc}(t)$ has a raised cosine spectrum that is zero for $|f| > \frac{1+\beta}{2T_o}$, and $H_{RF_i}(f)$, the Fourier transform of $h_{RF_i}(t)$, is zero only at discrete frequencies (of measure zero), it follows that there are an infinite number of impulse responses $h_{ci}(t) \neq h_{RF_i}(t)$ having a Fourier Transform $H_{ci}(f)$ equal to $H_{RF_i}(f)$ over $|f| < \frac{1+\beta}{2T_o}$. For the purpose of equalization, it does not matter which of these solutions is employed. For the i-th channel our approach yields a sample-spaced impulse response $h_{ci}[n]$ of length N_h ideally satisfying

$$\sum_{n=0}^{N_h-1} h_{ci}[n]e^{-j2\pi f n T_s} = \frac{1}{T_s} H_{RF_i}\left(\frac{f}{T_s}\right) \quad for \quad -\frac{1+\beta}{2T_o} < f < \frac{1+\beta}{2T_o} \tag{9}$$

In other words, if eqn.(8) holds and $h_{ci}(t) = \sum_{n=0}^{N_h-1} h_{ci}[n]\delta_a(t - nT_s)$, then

$$h_i(t) = \sum_{k=1}^{K} g_{ik} p_{rc}(t - \tau_k) = \sum_{n=0}^{N_h-1} h_{ci}[n] p_{rc}(t - nT_s) \tag{10}$$

over raised cosine bandwidth *even if* $\tau_k \neq nT_s$. The minimum number of taps required, N_h, and the time between samples, T_s, must roughly satisfy $(N_h - 1)T_s = \tau_{max}$, where τ_{max} is the multipath delay spread equal to the time duration of $h_{RF_i}(t)$. Assuming $\tau_{max} = 10\ \mu s$ and sampling every $T_s = 5\ \mu s$, our illustrative simulations presented in Section 4 use N_h=3 taps. Note that as the development to follow will show, we do not need to sample the output of each channel at the rate $f_s = \frac{1}{T_s}$, only the Nyquist pulse $p_{rc}(t)$ which is stored in memory. The blind identification algorithm to follow only requires two samples per symbol similar to classical fractionally spaced equalization.

For the sake of simplicity, assume that there are only $M = 2$ antennas. Consider symbol-spaced sampling at each of the two associated channels. From eqn. (6), it follows that

$$\begin{aligned} y_i[n] = y_i(nT_o) &= \sum_{m=0}^{N_b-1} s[m] h_i(nT_o - mT_o) \\ &= \sum_{m=0}^{N_b-1} s[m] h_i[n-m] \qquad i = 1,2 \end{aligned} \tag{11}$$

Assume eqn.(10) holds,

$$\begin{aligned} h_i[n] = h_i(nT_o) &= \sum_{k=0}^{N_h-1} h_{ci}[k] p_{rc}(nT_o - kT_s) \\ &= \sum_{k=0}^{N_h-1} h_{ci}[k] p_{rc}(nLT_s - kT_s) \\ &= \sum_{k=0}^{N_h-1} h_{ci}[k] p_{rc}[nL - k] \qquad i = 1,2 \end{aligned} \tag{12}$$

where $T_s = \frac{T_o}{L}$, where L is assumed to be an integer. $p_{rc}[n]$ is $p_{rc}(t)$ sampled at L times the symbol rate, i.e.,

$$p_{rc}[n] = p_{rc}(n\frac{T_o}{L}). \tag{13}$$

Again, it is only necessary to have $p_{rc}(t)$ sampled at the rate $f_s = \frac{L}{T_o}$ stored in memory. It is not necessary to sample either of the $M = 2$ channel outputs more than twice per symbol.

Whenever you have the same input into two different FIR filters, Xu et al[7] noted that:

$$y_i[n] * h_j[n] - y_j[n] * h_i[n] = 0 \tag{14}$$

Substituting the parametric model for either channel described by eqn.(12) yields

$$\left\{ \sum_{k=0}^{N_h-1} h_{c1}[k] p_{rc}[nL - k] \right\} * y_2[n] - \left\{ \sum_{k=0}^{N_h-1} h_{c2}[k] p_{rc}[nL - k] \right\} * y_1[n] = 0. \tag{15}$$

By the property of distributive law of convolution, this relationship can be rewritten as

$$\sum_{k=0}^{N_h-1} h_{c1}[k] \{p_{rc}[nL - k] * y_2[n]\} - \sum_{k=0}^{N_h-1} h_{c2}[k] \{p_{rc}[nL - k] * y_1[n]\} = 0 \tag{16}$$

Define

$$p_{rc}^{(k)}[n] = p_{rc}[nL - k] \qquad k = 0, 1, \ldots, N_h - 1 \tag{17}$$

and

$$y_i^{(k)}[n] = p_{rc}^{(k)}[n] * y_i[n] \qquad i = 1,2 \tag{18}$$

Then

$$\sum_{k=0}^{N_h-1} h_{c1}[k] y_2^{(k)}[n] - \sum_{k=0}^{N_h-1} h_{c2}[k] y_1^{(k)}[n] = 0 \qquad \forall n \tag{19}$$

The above equation can be expressed in a matrix form by defining the following quantities. First,

$$\mathbf{Y}_i = \begin{bmatrix} y_i^{(0)}[0] & y_i^{(1)}[0] & \cdots & y_i^{(N_h-1)}[0] \\ y_i^{(0)}[1] & y_i^{(1)}[1] & \cdots & y_i^{(N_h-1)}[1] \\ \vdots & \vdots & & \vdots \\ y_i^{(0)}[N_s - 1] & y_i^{(1)}[N_s - 1] & \cdots & y_i^{(N_h-1)}[N_s - 1] \end{bmatrix} \qquad i = 1,2 \tag{20}$$

where $N_s \leq N_b$ is the number of symbol-spaced samples used to estimate the channels. Further, define

$$\mathbf{h} = \left[\ h_{c1}[0],\ \cdots,\ h_{c1}[N_h-1],\ h_{c2}[0],\ \cdots,\ h_{c2}[N_h-1]\ \right]^T \tag{21}$$

Then, by eqn.(19),

$$\left[\mathbf{Y}_2 \vdots - \mathbf{Y}_1\right]\mathbf{h} = \mathbf{0} \tag{22}$$

Note that $\mathbf{h}$ can be solved by choosing the "smallest" right singular vector of $\left[\mathbf{Y}_2 \vdots - \mathbf{Y}_1\right]$.

We can improve the performance by generating additional equations using a different starting point, other than that at $t = 0$ corresponding to the peak of the first symbol for direct path. Assume symbol-spaced sampling of $y_i(t) = \sum_m s[m]h_i(t - mT_o)$ starting at $\frac{\ell}{L}T_o$ where $0 < \ell < L-1$. We here consider only the use of a single value of ℓ, *e.g.*, that value for which $\frac{\ell}{L} \approx \frac{1}{2}$ corresponding to symbol-spaced sampling from a starting point of $t = \frac{T_o}{2}$. Define

$$\begin{aligned} z_i[n] &= y_i(nT_o + \frac{\ell}{L}T_o) \\ &= \sum_{m=0}^{N_b-1} s[m]h_i(\frac{\ell}{L}T_o + nT_o - mT_o) \\ &= s[n] * g_i[n] \end{aligned} \tag{23}$$

where $g_i[n] = h_i(\frac{\ell}{L}T_o + nT_o)$. Then

$$\begin{aligned} g_i[n] &= \sum_{k=0}^{N_h-1} h_{ci}[k]p_{rc}(\frac{\ell}{L}T_o + nT_o - kT_s) \\ &= \sum_{k=0}^{N_h-1} h_{ci}[k]p_{rc}[\ell + nL - k] \qquad i = 1,2 \end{aligned} \tag{24}$$

Since $z_1[n] = s[n] * g_1[n]$ and $z_2[n] = s[n] * g_2[n]$, it follows that

$$z_1[n] * g_2[n] - z_2[n] * g_1[n] = 0 \tag{25}$$

Define

$$p_{rc_{off}}^{(k)}[n] = p_{rc}[\ell + nL - k] \qquad k = 0, \ldots, N_h - 1 \tag{26}$$

and

$$z_i^{(k)}[n] = p_{rc_{off}}^{(k)}[n] * z_i[n] \qquad i = 1,2 \tag{27}$$

A derivation similar to that done previously for a starting point of $t = 0$ leads to the following relationship for symbol-spaced sampling with a starting point of $t = \frac{\ell}{L}T_o$.

$$\sum_{k=0}^{N_h-1} h_{c1}[k]z_2^{(k)}[n] - \sum_{k=0}^{N_h-1} h_{c2}[k]z_1^{(k)}[n] = 0 \qquad \forall n \tag{28}$$

Recall $z_i[n] = y_i(\frac{\ell}{L}T_o + nT_o)$ and define

$$\mathbf{Z}_i = \begin{bmatrix} z_i^{(0)}[0] & z_i^{(1)}[0] & \cdots & z_i^{(N_h-1)}[0] \\ z_i^{(0)}[1] & z_i^{(1)}[1] & \cdots & z_i^{(N_h-1)}[1] \\ \vdots & \vdots & & \vdots \\ z_i^{(0)}[N_s-1] & z_i^{(1)}[N_s-1] & \cdots & z_i^{(N_h-1)}[N_s-1] \end{bmatrix} \tag{29}$$

Concatenating N_s equations in a fashion similar to that done in the case of symbol-spaced sampling starting at $t = 0$ yields

$$\left[\mathbf{Z}_2 \vdots - \mathbf{Z}_1\right] \mathbf{h} = \mathbf{0}. \tag{30}$$

Appending eqn.(30) to eqn.(22), this results in

$$\underbrace{\begin{bmatrix} \mathbf{Y}_2 & -\mathbf{Y}_1 \\ \mathbf{Z}_2 & -\mathbf{Z}_1 \end{bmatrix}}_{\mathbf{X}(N_s)} \mathbf{h} = \mathbf{0} \tag{31}$$

It is important to note that the **same** coefficients, $h_{ci}[k]$, characterize the channel $h_i[n]$, realized by symbol-spaced sampling starting at $t = 0$, **and** the virtual channel, $g_i[n]$, realized by symbol-spaced sampling starting at $t = \frac{\ell}{L}T_o$. This was not exploited in Xu's method[7]. Thus, in a sense, Xu et al's approach to virtual channels overparameterized the channel identification problem, leading to poorer performance and greater computational complexity than our method. This does not detract from the fact that Xu et al's contribution was very important and the underlying basis of our algorithm.

In the practical case where the channels are corrupted by noise, we follow the lead of Xu et al. and estimate $\mathbf{h}$ as that which minimizes $\|\mathbf{X}(N_s)\mathbf{h}\|^2$ subject to the constraint $\|\mathbf{h}\| = 1$. The solution is to choose $\mathbf{h}$ as that right singular vector of $\mathbf{X}(N_s)$ associated with the smallest singular value.

4 Simulations

In the simulations, the multipath signals arrived at -5^o, 0^o, and 5^o with complex amplitudes of -0.8, 1, 0.9j, respectively, and time delays of 10, 0, and 5 micro-seconds. The array comprised of M=2 antennas separated by three wavelengths. The symbol alphabet was QPSK. The symbol period was 25 micro-seconds, i.e., the symbol rate was 40 KHz. The symbol waveform was a pulse with a square-root raised cosine spectrum with β=0.35. The baseband bandwidth was 27 KHz (the RF bandwidth was 54 KHz.)

We compared the new approach developed in the previous section with Xu et al's method [7] in the simulated environment described above. Since $\tau_{max} = 10\ \mu s$ and setting $T_s = 5\ \mu s$, the new method required N_h=3 taps to characterize each of the two channels. For Xu et al's method, since τ_{max} was a fraction of the duration of $p_{rc}(t)$, the number of taps needed is roughly the number of symbol times, T_o, encompassed by the effective duration of $p_{rc}(t)$. For β=0.35, Xu et al's method required 10 coefficients per channel.

For the first set of results presented in Figure 1, each of the two respective algorithms were applied to a block of $N_s = 30$ symbols. The results of ten independent trials runs for each of the two algorithms are overlaid in Figures 1(a) to 1(d). Note that only the magnitude of the channel estimates are plotted; similar performance was observed with respect to estimating the phase of each channel. Figures 1(a) and 1(b) correspond to channels 1 and 2, respectively, at an SNR of 35 dB, while Figure 1(c) and 1(d) correspond to channels 1 and 2, respectively, at an SNR of 45 dB. In each of these plots, the solid curve represents the magnitude of the actual channel, $h_i[n] = \{h_{RF_i}(t) * p_{rc}(t)\} |_{t=nT_o}$, $i = 1, 2$. The dash-dot curves represent the magnitude of the channel estimates obtained by the new algorithm, while the dashed curves represent the magnitude of the channel estimates obtained by Xu et al's algorithm. The new algorithm is observed to provide channel estimates that very closely match the respective magnitudes of the true channels in all runs. In contrast, even at a very high SNR of 45 dB, Xu et al's original method is unable to identify the channels given a block of only $N_s = 30$ symbols.

For the second set of results presented in Figure 2, Xu et al's original method was applied to a block of $N_s = 200$ symbols. The new method was still applied to a block of only $N_s = 30$ symbols. The result of only one trial run for each of the two algorithms is plotted in Figures 2(a) to 2(f). Despite still only being supplied with $N_s = 30$ symbols, the new algorithm is observed to perform quite well even when the SNR is lowered to 25 dB. In contrast, despite increasing the block length from $N_s = 30$ to $N_s = 200$ symbols, Xu et al's algorithm is not able to accurately identify the channels until the SNR is increased to 45 dB.

References

[1] Lang Tong, Guanghan Xu, and Thomas Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach", *IEEE Transactions on Information Theory*, vol. 40, pp. 340–349, March 1994.

[2] Eric Moulines, Pierre Duhamel, J-F Cardoso, and Sylvie Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters", *IEEE Transactions on Signal Processing*, vol. 43, pp. 516–525, February 1995.

[3] Y. Li and Z. Ding, "New results on the blind identification of fir channels based on second order statistics", *IEEE Milcom '93,* pp. 644–647, October 1993.

[4] Jitendra K. Tugnait, "Blind equalization and estimation of FIR communications channels using fractional sampling", *IEEE Transactions on Communications*, vol. 44, pp. 324–336, March 1996.

[5] D.H. Brooks and C.L. Nikias, "Multichannel adaptive blind deconvolution using the complex cepstrum of higher-order cross-spectra", *IEEE Transactions on Signal Processing*, vol. 41, pp. 2928–2934, September 1993.

[6] D.N. Godard, "Self-recovering equalization and carrier tracking in two- dimensional data communication system", *IEEE Transactions on Communications*, vol. 28, pp. 1867–1875, November 1980.

[7] G. Xu, Hui Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification", *IEEE Transactions on Signal Processing*, vol. 43, pp. 2982–2993, December 1995.

[8] Timothy A. Thomas and Michael D. Zoltowski, "Novel receiver signal processing for interference cancellation and equalization in cellular TDMA communication", *in IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, pp. 3881–3884, April 1997.

[9] Timothy A. Thomas and Michael D. Zoltowski, "Novel receiver space-time processing for interference cancellation and equalization in narrowband TDMA communication", *in IEEE Int'l Conf. on Vehicular Technology*, pp. 160–164, May 1997.

[10] Timothy A. Thomas and Michael D. Zoltowski, "Nonparametric channel identification and multichannel equalization for narrowband digital communications", *IEEE Milcom '97*, 2-5 Nov. 1997.

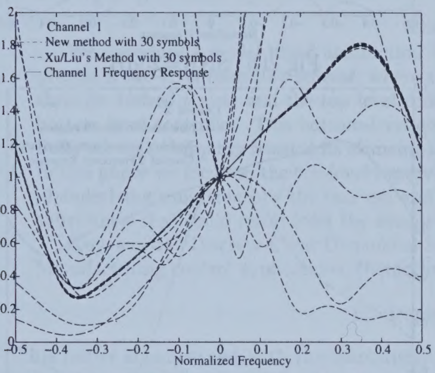


Fig.1(a) SNR=35dB, 30 symbol block

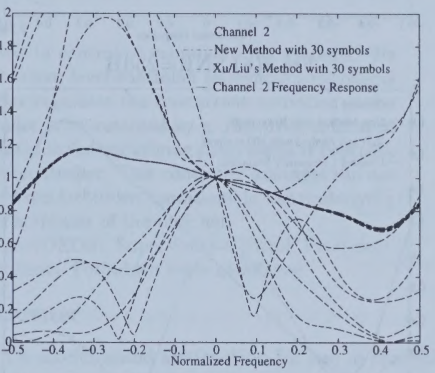


Fig.1(b) SNR=35dB, 30 symbol block

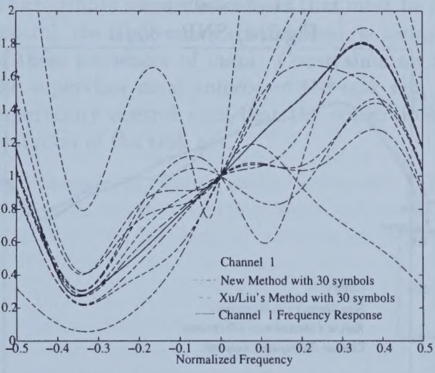


Fig.1(c) SNR=45dB, 30 symbol block

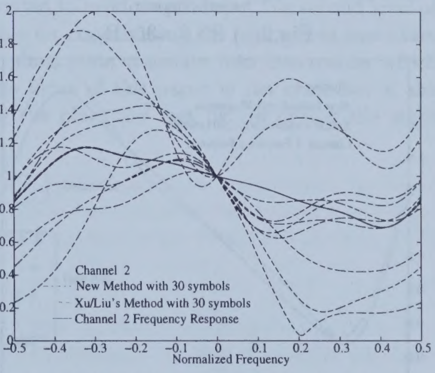


Fig.1(d) SNR=45dB, 30 symbol block

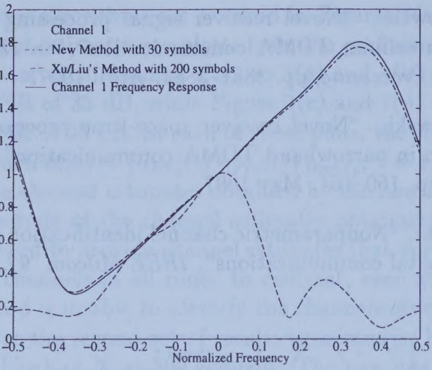


Fig.2(a) SNR=25dB

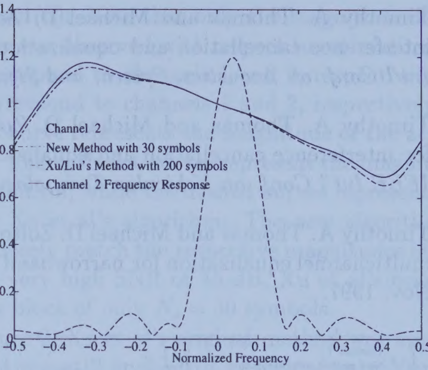


Fig.2(b) SNR=25dB

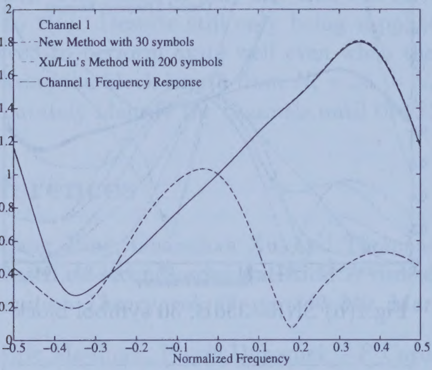


Fig.2(c) SNR=35dB

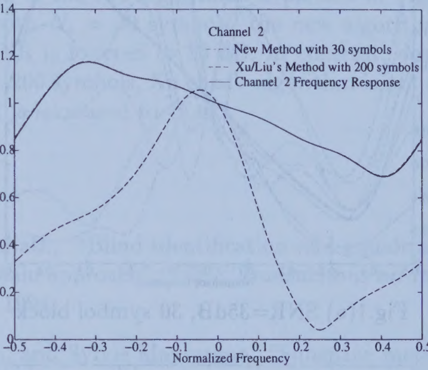


Fig.2(d) SNR=35dB

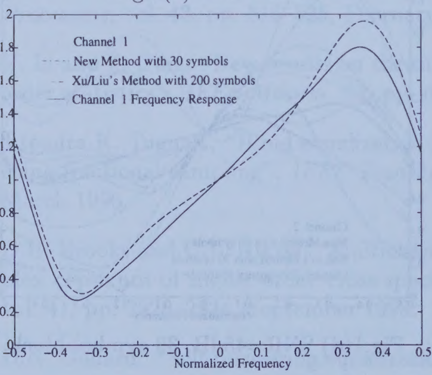


Fig.2(e) SNR=45dB

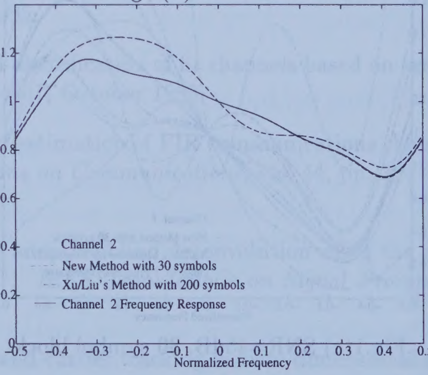


Fig.2(f) SNR=45dB

Supervisory Control of Concurrent Task Sequences Operating Over Low-Level Devices

Jeffrey Ashley, Jr. and Lawrence E. Holloway *
Department of Electrical Engineering and
Center for Robotics and Manufacturing Systems
University of Kentucky
Lexington, KY 40506

Abstract

We introduce a model based abstraction for use in control of manufacturing systems. We present a two-level controller scheme, where the bottom level controller sequences subsystems through desired states, and the top level controller regulates the concurrent activities of the bottom-level controller. This bottom-level controller is represented by a *Task Net*, a form of Petri net with mappings relating its elements to states in the underlying subsystems it controls. In this paper we focus on the top-level supervisory controller. This controller manages the use of underlying subsystems by the task net and prevents forbidden conditions in the underlying subsystems that could result from the concurrent activities of the task net.

Key words:- Discrete Event Dynamical Systems(DEDS), Supervisory Control, Petri nets, Manufacturing control applications, Hierarchal systems, Forbidden state avoidance.

I. INTRODUCTION

This paper is concerned with the introduction of a model based abstraction for use in control of manufacturing systems. We consider a two-level control scheme as illustrated in figure 1. The bottom level control interacts directly with the plant. It is represented as a *task net*, which defines sequences of tasks that must be executed to produce product. The second level of control, the supervisory control level, is responsible for coordinating the concurrent execution of these sequences of tasks. Constraints on the plant state translate into constraints which the supervisor must enforce on the task net. The focus of this paper is the synthesis of the supervisory control such that the constraints on the plant are met by restricting the state evolution of the task net.

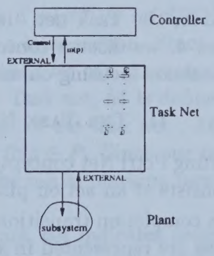


Fig. 1. The Controller structure

The plant $\mathcal{G}$ is the manufacturing system to be controlled. The bottom layer of the controller is referred to as a *task net*, and is an abstraction of a control program. The task net is composed of sequences of *tasks* (defined formally in the next section), which represent portions of the control program which interacts with the plant. Such tasks may be automatically synthesized by analysis of the plant model, as in [1].

---

*Please address all correspondence to Jeff Ashley or L. E. Holloway at the above address. Phone: (606) 257-6262 ext. 203. email: jashl01@engr.uky.edu or holloway@engr.uky.edu. The first author is a doctoral candidate under the tutelage of the second author. The second author has been supported by Rockwell International, NSF grant ECS-9308737, NASA grant NGT-40049, ARO grant DAAH04-96-1-0399, and the Center for Robotics and Manufacturing Systems at the University of Kentucky.

The task net is represented as a Petri net, and each task is a short sequence of places and transitions in the net. The manner the tasks are connected together in the net indicates the sequence of operations that must be undertaken to produce a part, and thus indicates the structure of the control code. The marking (state) of the task net represents which portions of the control program are active at a particular time. Multiple portions of the program may be concurrently active, i.e. there may be several tasks in the task net which could be marked simultaneously.

The top-level controller is the *supervisory logic controller* (SLC). The supervisory logic controller is responsible for ensuring that multiple tasks don't try to simultaneously access a resource, and for ensuring that the underlying system does not reach any forbidden states. The controller does these by regulating the initiation of tasks, in effect by restricting the flow of tokens in the task net. This means the supervisor must have information about how each task involves the states of the underlying subsystems.

There is a considerable body of work on supervisory control of discrete event systems (e.g. [2], [3], [4], [5]). Typically such work considers only a single level of control - the supervisor must directly enforce sequence constraints or state constraints specified on the plant. In our work, there is an intermediate controller (the task net), which represents the sequences of tasks necessary to produce a product, and which contains specific control code for interacting with the plant. In our work, the supervisor enforces the constraints on the plant through the regulation of activities in this intermediate controller. This is similar to the work of Tittus, Fabian, and Lennartson [7] for the coordination of recipes in batch applications. Another interesting work is [6]. In [6], Zhong and Wonham investigate hierarchical supervisors and identify some critical requirements for these types of supervisors schemes.

The task net may be synthesized through analysis of a plant model, as in [1]. Using this plant model, we associate substates of the plant to the tasks in the task net. This plant substate information is used for managing resources and preventing the plant from reaching forbidden markings. By working at the level of the task net, we can efficiently handle simultaneously executing sequential logics without having to analyze the potentially much larger plant model. The supervisory controller then operates on the task net in a state feedback control similar to [8; 9].

In the next section, we introduce the Task net. In section 3, we present our supervisory controller mechanism. In section 4, we discuss controller synthesis and present our results. Finally, we close with a brief discussion focusing on areas of future research.

II. THE TASK NET

In this section we build on existing Petri Net concepts to develop the notion of a *task* and the *Task Net System*. A task, $\mathcal{T}$, consists of an action place, $p_{\mathcal{T}}^{act}$, a maintenance place, $p_{\mathcal{T}}^{main}$, an initiation transition, $t_{\mathcal{T}}^{begin}$, and a completion transition, $t_{\mathcal{T}}^{comp}$ connected as illustrated in figure 2. In the standard manner, places are represented in a Task net as circles, and transitions are represented as bars. The set of all tasks is defined as $SetOfTasks$.

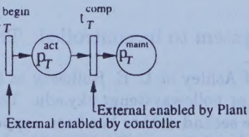


Fig. 2. A Task

A generic definition of a task in the Task net model would be: A transition-place-transition-place model that represents some sequence of event occurrences and state observations in the Plant. We call this sequence a *task sequence*. So, a task represents some interaction between the controller and the plant. When the first place, $p_{\mathcal{T}}^{act}$, is marked in a task sequence, the action is begun by the controller and the plant. This may lead the plant through several states. When the transition $t_{\mathcal{T}}^{comp}$ fires and $p_{\mathcal{T}}^{main}$ is marked (see figure 2), then the action has been completed and the manufacturing system is ready for the next task. The initiation transition $t_{\mathcal{T}}^{begin}$ for each task is externally controlled in the sense that a supervisor may enable or disable the initiation of the task. The completion transition $t_{\mathcal{T}}^{comp}$ is not externally controlled, so that once the task is begun, it will continue until completion. Thus, $t_{\mathcal{T}}^{comp}$ often will represent a sensing or other acknowledgement from the plant that the task was completed.

A task represents some localized action. A sequencing of tasks necessary for product flow or resource usage is represented as a Task Net.

A Task Net is a tuple:

$$S = (P, T_c, T_u, Pre, Post)$$

Where P is the set of places, T_c is the set of *controlled transitions*, T_u is the set of *uncontrolled transitions*, *Pre* and *Post* are directed arcs between places and transitions. P is the set of all places for all tasks, $P = \{p_{\mathcal{T}}^{act}, p_{\mathcal{T}}^{main} \mid \mathcal{T} \in SetOfTasks\}$. Likewise, T_u and T_c are defined by: $T_u = \{t_{\mathcal{T}}^{comp} \mid \mathcal{T} \in SetOfTasks\}$; $T_c = \{t_{\mathcal{T}}^{begin} \mid \mathcal{T} \in SetOfTasks\} \cup T_e$, where T_e is a set of *exit transitions.*

The following restricts the type of Task net models that we consider for this paper. Generalizing the models we consider is a subject of future research.

Restriction 1 : We require that our task net be composed of linear sequences of tasks. Each place and transition will have at most one arc leading into and out of it.

These sequences are called *task net strings*. Let a task net string be denoted as $\mathcal{TS}$. The set of all task net strings is $SetOf\mathcal{TS}$. We must also append an *exit transition* to each task net string in the Task Net. So a $\mathcal{TS}$ is merely an ordered sequence of some number of tasks.

The state of the system is described by the *marking function*. The marking $m(p)$ is defined as the number of tokens in place p. We say a place is *marked* if there exists a token in that place. As a prelude to the following we must mention that our controller will, among other things, enforce safeness in our Task net model. We define *task safeness* in a manner similar to safeness in Petri net literature [10]. A marking is task safe if $m(p_{\mathcal{T}}^{act}) + m(p_{\mathcal{T}}^{main}) \leq 1$, $\forall \mathcal{T} \in SetOfTasks$. The enforcement of task safeness is covered in a later section. Since no place will contain more than one token when the system is task safe, then we can characterize the states of the net by the set of marked places: the marking of the Task net, M is defined as the set of all places, $p \in P$ such that $m(p) \geq 1$. The *initial marking* of the task net , $M_{INIT} = \emptyset$, is defined such that each place contains no token, or $m(p) = 0 \forall p \in P$. Since our task net represents product flow and resource usage for the real system, the initial marking merely implies that no tasks are in progress upon startup.

The evolution of the marking of the task net indicates progress in the desired sequence of operations for the plant. The marking evolves by the firing of transitions in the net. Transitions may only fire if they are both *state enabled* and *externally enabled*. A transition is *state enabled* if the place with an arc leading into it contains a token, i.e. $m(p) \geq 1$ for the place such that p has an arc into the transition. If a transition does not have a place with an arc leading into it we assume it is always state enabled. Each transition t in the task net has a boolean function $EXTERNAL : t \rightarrow \{TRUE, FALSE\}$. A transition t is *externally enabled* if the function $EXTERNAL(t) = TRUE$. For task initiation transitions, $t \in T_c$, external enabling corresponds to the permission to commence a task. For the task completion transitions, $t \in T_u$, the external enabling represents the sensing or acknowledgement by the system that the task has been completed. Any transition that is state enabled and externally enabled in a task net can *fire*. thus removing a token from the place with an arc into the transition and adding a token to the place with an arc leading from the transition. We assume that only one enabled transition can fire at any given time.

We model our system resources(e.g. the actual machinery) as simple Petri net models, then use these models to generate the task sequences. If this is the case then specific state information about the system will be readily available. The system which we intend to control is modeled by a set of discrete state subsystems, $\mathcal{G} = \{G_1, G_2, \ldots G_n\}$. The dynamics of the subsystems are beyond the scope of this paper. However, for each subsystem G_i, let $\mathcal{X}_{G_i}$ denote the set of states that G_i could have. Define $X_{init,G_i} \subset \mathcal{X}_{G_i}$ be the initial state of the subsystem. and define $X_{init,\mathcal{G}} = \cup_{1 \leq i \leq n}\{X_{init,G_i}\}$ as the initial state of the entire system. Let $\mathcal{X}_{\mathcal{G}}$ denote the set of all possible states of the system, $\mathcal{X}_{\mathcal{G}} = \cup_{1 \leq i \leq n}\{\mathcal{X}_{G_i}\}$. [1]

Each task in the task net interacts with the plant through some underlying sequence of actions. However, in this paper, we are only concerned about the effects of these actions on the states of the subsystems and their dependence on the states of the subsystems. We assume that each task only affects a single system, and we denote this system as $G(\mathcal{T})$ for task $\mathcal{T}$. Our goal is to eliminate the need to consider the specific details of the actual task sequences in our envisioned system. We are now ready to define the mapping $PtoX$. $PtoX$ defines the relationship between task net places and the condition-event system marking.

Definition 1 : Given a task $\mathcal{T}$, define the mapping $PtoX : P \to 2^X$ such that $PtoX(p_{\mathcal{T}}^{act}) \subseteq \mathcal{X}_{G(\mathcal{T})}$ is the set of states that subsystem $G(\mathcal{T})$ passes through while $p_{\mathcal{T}}^{act}$ is marked. and $PtoX(p_{\mathcal{T}}^{main}) \subseteq \mathcal{X}_{G(\mathcal{T})}$ is the set of states that subsystem $G(\mathcal{T})$ is in while $p_{\mathcal{T}}^{main}$ is marked.

$PtoX$ takes advantage of the fact that whenever an action place is marked there is an underlying sequence of transitions that fire in some subsystem model G_i. These transition firings lead to a deterministic progression of the marking that can be merely lumped into one set.

From the mapping $PtoX$, we derive two other mappings: $PtoSS$, and $TtoX$. The first, $PtoSS : P \to G_i$, maps each place in the task net to its appropriate subsystem, G_i. If a place p is part of task $\mathcal{T}$, then $PtoSS(p) = G(\mathcal{T})$. Define $TtoX : T_C \to 2^X$ such that if the firing of $t_{\mathcal{T}}^{begin}$ leads to $p_{\mathcal{T}}^{act}$ being immediately marked then $TtoX(t_{\mathcal{T}}^{begin}) = PtoX(p_{\mathcal{T}}^{act})$. Inverse mappings are defined in the obvious way. The inverse mapping of $PtoX$ is $PtoX^{-1}(x) = \{p \mid p\, s.t.\, x \in PtoX(p)\}$. $PtoX^{-1}$ takes as an argument a state x in the condition-event system and returns the set of places in the task net that it occurs. Likewise, $PtoSS^{-1} : G_i \to 2^P$, maps from each subsystem, G_i, to a set of places in the Task Net.

We impose the following restriction on tasks and condition-event systems.

Restriction 2 :

1. Before a task is *initiated* (i.e. its controllable transition fires) the subsystem must be in the proper initial state. We denote this as $X^{\mathcal{T}}_{init,G_i}$. $X^{\mathcal{T}}_{init,G_i}$ for a task corresponds to an initial marking for all elemental nets associated with the subsystem, G_i.
2. When a maintain-place is marked, the equivalent condition-event system G_i will be in a *non-evolving* state. By non-evolving we mean that $PtoX$ maps to a single marking.

Item 1 requires that the system be in a known state at the start of a task. Item 2 requires that our system be non changing (in terms of event occurrences in G_i) at the completion of a task. See figure 3.

A $PtoX$ mapping for a maintain-place would correspond to the final state of a subsystem model G_i after a task had been executed. An equivalent state for an action place would contain the set of places marked in the subsystem model G_i during the execution of a task sequence.

Knowledge of the initial state of the Plant and its underlying condition-event system is

---

[1] We use noncalligraphic X to indicate sets of places that represent single markings within the system $\mathcal{G}$ or subsystem G_i of interest. We use the calligraphic $\mathcal{X}$ to indicate sets of places that represent one or more markings within a subsystem or within the system.

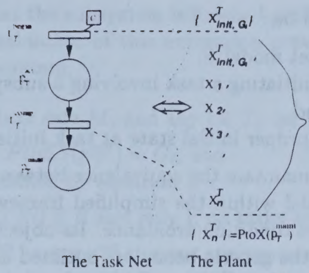


Fig. 3. Task interpretation

required for efficient controller synthesis. The *Task net property* requires that this initial state is preserved in the Task net sequences.

Definition 2 : A Task Net S in its initial state and a condition-event system $\mathcal{G}$ satisfies the *TN property-A*, if we can externally enable all controllable transitions in S, send a single token through the Task net, and the following holds true:

1. For each task that is initiated, the appropriate subsystem will be in its initial state for that task, X^T_{init,G_i}, prior to task initiation.
2. After the token leaves the Task Net, $\mathcal{G}$ will be back in the state $X_{init,\mathcal{G}}$.

The Task Net property-A ensures that the system can have at least one token in the Task net at any one time . We will require one more property of the Task net, defined in the next section.

III. THE SUPERVISORY LOGICAL CONTROLLER

The supervisory logic controller (SLC) is a simple state feedback controller that takes the marking of the Task Net as its input. The SLC evaluates its set of expressions for the current marking and determines a set of controllable transitions that should be disabled. All controllable transitions are assumed to be external enabled unless our controller disables one or more of them. Since places following these transitions initiate a state evolution in the plant, we can indirectly control the plant using a state feedback policy via the firing of these transitions. Our controller uses expressions to achieve a set of objectives which include: task safeness; shared resource management; and forbidden state avoidance.

Each expression we consider is represented as a set of places, $P' \subseteq P$. We use a *boolean valuation* $\overline{\Delta}$ to evaluate these expressions as true or false for a given marking. Given a marking M, and some set of places $P' \subseteq P$, define the evaluation of P' as follows:

$$\overline{\Delta}(P') = TRUE \text{ iff } m(p) = 0 \forall\, p \in P'$$

The supervisory controller associates an expression with the initiation transition for each task, i.e. with each $t \in T_c$. The supervisory controller disables the transition if the expression evaluates to FALSE under the current marking, and externally enables the transition if the expression evaluates to TRUE. In the next section, we present methods for constructing the expressions of these transitions such that the supervised Task net and plant satisfy desired properties.

IV. EXPRESSION CONSTRUCTION

We present a method of determining the set of expressions that our controller uses for decision making. In this section, we cover the generation of four types of expressions, three for resource objectives and one for forbidden condition avoidance.

We consider resource objectives to be :

1. Ensuring safeness in the Task net model S.
2. Preventing the controller from initiating a task involving a subsystem when the subsystem is already performing a task; and
3. Ensuring a subsystem is in the proper initial state at task initiation.

These resource objectives exist to maintain the equivalence between the Plant, the condition-event system, and the task net model within the simplified framework we present. The last type of expression is one for forbidden condition avoidance. Its objectives are outlined as formal specifications as distinguished from the generic behaviors outlined above.

We will use the $\overline{\Delta}$ function to ensure these desired behaviors occur. We present these first since the theoretical results build upon each other in this order. We treat these objectives separately and develop expressions for each.

A. Safeness expressions

The first duty of the supervisory logic controller is the prevention of markings that do not satisfy the definition of task safeness. Instead of structurally guaranteeing safeness within the Task net, we enforce Task safeness using the controller. This simplifies the model by eliminating the cycles needed to ensure safeness, and it does not add much in the way of added work to the controller.

We present a simple expression construction method that generates an expression for each transition that can be used by $\overline{\Delta}$ to guarantee task safeness. Let $Expr_{safe}(t)$ be the expression that is used to ensure task safeness in the Task net model. For each $\mathcal{T} \in SetOfTasks$, define $Expr_{safe}(t_{\mathcal{T}}^{begin})$ using the following:

$$Expr_{safe}(t_{\mathcal{T}}^{begin}) = \{p_{\mathcal{T}}^{act}, p_{\mathcal{T}}^{main}\} \tag{1}$$

$Expr_{safe}(t)$ contains the two places immediately following the controllable transition t. The following lemma states that we ensure task safeness in our Task net model.

Lemma 1 *:* Given a Task net S and some task safe marking M, if each $t \in T_C$ is enabled to fire by the supervisor only when $\overline{\Delta}(Expr_{safe}(t)) = TRUE$, then all subsequent markings will also be task safe.

Proof: Each transition $t \in T_C$ has an $Expr_{safe}(t)$ that contains the action and maintenance places immediately following the transition. If either of these places is marked then $\overline{\Delta}(Expr_{safe}(t)) = FALSE$. This will externally disable transition t thereby preventing a token from entering either two places if either is marked. Thus, from any task safe marking, all next markings will be task safe also, so all subsequent markings are task safe.

□□□

B. Shared Resource expressions

These expressions are concerned with ensuring that a subsystem is never simultaneously utilized by two or more tasks. In other words, we use these types of expressions to guarantee mutually exclusive use of each subsystem.

Let $Expr_{shared}(t)$ be the expression set that is used to ensure that a subsystem is never used by more than one task at any time. For all $t \in T_C$, we develop an $Expr_{shared}(t)$ using the following:

$$Expr_{shared}(t_{\mathcal{T}}^{begin}) = \{p' \mid PtoSS(p_{\mathcal{T}}^{act}) = PtoSS(p') \; AND \; p' \equiv p_{\mathcal{T}'}^{act} \; for \; some \; \mathcal{T}'\} \tag{2}$$

Thus, we disable $t_{\mathcal{T}}^{begin}$ if any other action place in the task net associated with the same subsystem is marked. We only include action places in $Expr_{shared}(t)$ because we assume that once a $p_{\mathcal{T}}^{main}$ place is marked that the subsystem is finished performing a task. The following lemma shows directly that our evaluation of this expression prevents two tasks from simultaneously attempting to use one subsystem.

Lemma 2 : Given S, a task safe M, and any $t \in T_C$ let $\mathcal{T}$ be the task such that $t = t_{\mathcal{T}}^{begin}$. If

1. For this action place: $PtoSS(p_{\mathcal{T}}^{act}) = G_i$, and
2. $Expr_{shared}(t)$ is determined by equation 2.

Then $\overline{\Delta}(Expr_{shared}(t)) = TRUE$ if and only if no tasks are using the subsystem G_i.

Proof: only if- $PtoSS(p_{\mathcal{T}}^{act})$ maps $p_{\mathcal{T}}^{act}$ to some appropriate subsystem, G_i. $PtoSS^{-1}(G_i)$ maps G_i to all action places associated with G_i, and $Expr_{shared}(t)$ would include any such places. Therefore if any other task is using a subsystem then $m(p) \geq 1$ for some $p \in Expr_{shared}(t)$ and $\overline{\Delta}(Expr_{shared}(t))$ would be set to $FALSE$. This would disable the transition, t.

if- Using similar logic to above. If no other task $\mathcal{T}$ is using subsystem G_i then no action place such that $p \in PtoSS^{-1}(G_i)$ is marked. Then $\overline{\Delta}(Expr_{shared}(t))$ would be equal to $TRUE$ thereby externally enabling t. □□□

Or, we can prevent the next marking from violating our shared resource requirement provided we start with a marking that does not violate this requirement. It follows from the previous lemma and the empty initial marking(M_{INIT}) of the Task net that our controller will prevent any resource from being simultaneously used by more than one task.

C. Initial state expressions

We must also ensure that any subsystem that is about to perform a task is in the proper initial state, $X_{init,G_i}^{\mathcal{T}}$. We call the expressions that handle this issue *initial state expressions* and their generation is somewhat more involved. Such an expression is labeled by $Expr_{init}(t)$. For each $t \in T_C$, we use the $PtoX$ mapping and the Task net property to determine what places in the Task net have a subsystem state such that it is not equal to $X_{init,G_i}^{\mathcal{T}}$. We know then that if any of these places are marked the subsystem is not in $X_{init,G_i}^{\mathcal{T}}$, and so we merely append these places to $Expr_{init}(t)$.

First we use the $PtoX$ mappings for each place and $X_{init,\mathcal{G}}$ to build condition-event marking sets for each place p for each task net string. We use the algorithm shown in figure 4 to build marking sets for $\mathcal{G}$ for every place in a task net string. We denote the marking set over $\mathcal{G}$ associated with a place in the Task net as $\mathcal{X}_{\mathcal{G}}(p)$. Note that $\mathcal{X}_{\mathcal{G}}(p_{\mathcal{T}}^{act})$ can contain more than one marking for G_i while $\mathcal{X}_{\mathcal{G}}(p_{\mathcal{T}}^{main})$ represents a single marking.

Fig. 4.

Procedure: $\mathcal{X}_{\mathcal{G}}(p)$ construction :

```
For each $\mathcal{TS} = (\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3\cdots\mathcal{T}_n)$ in $SetOf\mathcal{TS}$ {
  $X' \Leftarrow X_{init,\mathcal{G}}$;
  For $i = 1$ to $n$ {
    $\mathcal{X}_{\mathcal{G}}(p_{\mathcal{T}_i}^{act}) \Leftarrow X' \cup PtoX(p_{\mathcal{T}_i}^{act})$;
    $\mathcal{X}_{\mathcal{G}}(p_{\mathcal{T}_i}^{main}) \Leftarrow (\mathcal{X}_{\mathcal{G}}(p_{\mathcal{T}_i}^{act}) - PtoX(p_{\mathcal{T}_i}^{act})) \cup PtoX(p_{\mathcal{T}_i}^{main})$;
    $X' \Leftarrow \mathcal{X}_{\mathcal{G}}(p_{\mathcal{T}_i}^{main})$ ;
}}}
```

We then use the procedure outlined in figure 5 to determine our $Expr_{init}$ for each transition in T_C.

For the following proof we define the following. Let t's firing lead to a task such that its required initial state is $X_{init,G_i}^{\mathcal{T}}$.

Lemma 3 : Given a Task net S and a task safe marking M such that $\overline{\Delta}(Expr_{shared}(t)) = TRUE \forall t \in T_C$. then $\forall t \in T_C, Expr_{init}(t)$ and the valuation $\overline{\Delta}(Expr_{init}(t))$ for these expressions ensure a task is initiated only if the subsystem is in the correct initial state, $X^{\mathcal{T}}_{init,G_i}$.

Proof: We have initial marking, and we build using the *PtoX* mappings in figure 4, the marking of the condition-event system at every place for all the Task net strings. Any places such that the marking of the subsystem G_i is not equal to $X^{\mathcal{T}}_{init,G_i}$ are incorporated into $Expr_{init}(t)$. So if any of these places are marked then $\overline{\Delta}(Expr_{init}(t)) = FALSE$. Therefore we only external enable a transition t whenever the subsystem is in the proper initial state for that task. □□□

Fig. 5.

```
Procedure: $Expr_{init}$ construction :
For each task $\mathcal{T}$ in $SetOfTasks$ {
    $Expr_{init}(t^{begin}_{\mathcal{T}}) \Leftarrow \emptyset$ ;
    For all $\mathcal{T}' \neq \mathcal{T}${
        if ( $X^{\mathcal{T}}_{init,G_i} \not\subseteq \mathcal{X}_G(p^{main}_{\mathcal{T}'})$ ) {
            $Expr_{init}(t^{begin}_{\mathcal{T}}) \Leftarrow Expr_{init}(t^{begin}_{\mathcal{T}}) \cup \{p^{act}_{\mathcal{T}'}, p^{main}_{\mathcal{T}'}\}$
} } }
```

It is interesting though rather trivial to note the following. If each task for a subsystem has the same $X^{\mathcal{T}}_{init,G_i}$ condition and the *PtoX* mapping for each of these tasks' maintain-place equals $X^{\mathcal{T}}_{init,G_i}$, then $Expr_{init}(t)$ will create minimal delays in the availability of subsystems between executing different tasks. This is because upon completion of one task, the subsystem will be in a state that would allow it to immediately execute a new task. In the real plant, that would correspond to a subsystem starting all tasks from one initial condition(idle) and ending up back at that state at the completion of a task.

D. Forbidden Condition expressions

The next issue becomes the forbidden condition problem across the interacting subsystems in our model. We specify forbidden condition pairs in terms of elemental net states. These do not require specific knowledge of the task net, and represent more fundamental states of the system.

A *forbidden condition* , F is specified as a pair of states, (x_i, x_j) in the plant. Given a forbidden condition, we wish to prevent its occurrence in the real system. Define the set of all forbidden condition pairs to be $\mathcal{F} \subseteq 2^X$.

We say a marking, M, is *forbidden condition free*, if under $\mathcal{X}_G(p), \mathcal{X}_G(p')$ for any $p, p' \in M$, there does not exist any $x_i \in \mathcal{X}_G(p)$, $x_j \in \mathcal{X}_G(p')$ such that $(x_i, x_j) \in \mathcal{F}$. A *Forbidden condition avoiding policy* for all forbidden condition pairs $(x_i, x_j) \in \mathcal{F}$ is an external enabling assignment for each $t \in T_c$ such that, given a forbidden condition free M, the next marking M' will also be forbidden condition free.

Definition 3 : A Task Net S in its initial state and a plant $\mathcal{G}$ satisfies the *TN property-B*, if we can externally enable all controllable transitions in S , send a single token through the Task net and each marking will be *forbidden condition free.*

TN property-B ensures that our *task net strings* do not achieve any forbidden conditions whenever they are executed singly. This is merely an extension of our original task net property to the notion of forbidden state avoidance. TN property-B property also ensures that tasks do not reach a forbidden state when executed singly. These problems are assumed to have been previously addressed. This partitions the forbidden state avoidance problem in an appropriate manner for our envisioned system.

We define the expression set , $Expr_{FS}$, used for forbidden state avoidance with the following:

$$Expr_{FS}(t_{\mathcal{T}}^{begin}) = \{p' \mid \exists x_j \in \mathcal{X}_{\mathcal{G}}(p') \; and \; x_i \in TtoX(t_{\mathcal{T}}^{begin}) \; such \; that \; (x_i, x_j) \in \mathcal{F}\} \qquad (3)$$

We show with the following lemma, that $Expr_{FS}$ when evaluated for all controllable transitions will be forbidden condition avoiding. We note here that this policy does not ensure that our system does not deadlock.

Lemma 4 *:* Given a Task net S and a task safe, forbidden condition free marking M, then $Expr_{FS}(t) \forall t \in T_C$ and the valuation $\overline{\Delta}(Expr_{FS}(t))$ over these transitions is forbidden condition avoiding.

Proof: Since, our current marking is forbidden condition free and since $PtoX(p_{\mathcal{T}}^{main}) \subseteq PtoX(p_{\mathcal{T}}^{act}), \forall \mathcal{T} \in SetOfTasks$, we know that all reachable markings under the most restrictive control(i.e. all controllable transitions disabled) are forbidden condition free. Now, let's consider $Expr_{FS}(t)$ for any arbitrary $t \in T_C$ to see if $\overline{\Delta}(Expr_{FS}(t))$ is forbidden condition avoiding for all $(x_i, x_j) \in \mathcal{F}$ such that x_i is in $TtoX(t)$. Note that the first element in the forbidden condition is chosen for illustration reasons and that it could just as easily be x_j.

From equation 3, we immediately see that any places $p \in P$ such that $x_j \in \mathcal{X}_{\mathcal{G}}(p)$ are included in $Expr_{FS}(t)$. If any of these places are marked in M then $\overline{\Delta}(Expr_{FS}(t))$ would be set to $FALSE$, thereby disabling the transition. It follows that under M, any $t \in T_c$ such that t's firing could lead to a marking that is not forbidden condition free will be externally disabled by $\overline{\Delta}(Expr_{FS}(t))$. Since only transitions in T_c that lead to forbidden condition free marking are externally enabled then $\overline{\Delta}(Expr_{FS}(t)) \forall t \in T_c$ is forbidden condition avoiding, □□□

E. The SLC and Deadlock

We have shown that given a task net and marking such that the system is: task safe, has no subsystems in $\mathcal{G}$ that are used by more than one task, and violates no forbidden conditions, then we can ensure all of the properties mentioned in this paper for the next marking.

Our controller is of the disabling type and our $Expr$ expressions represent disabling conditions, so the unioning of these sets also serves to disable when required. Given the empty initial marking for a Task net $M_{INIT} = \emptyset$ and $Expr = Expr_{safe} \cup Expr_{shared} \cup Expr_{init} \cup Expr_{FS}$, $\overline{\Delta}(Expr(t))$ for all controllable transitions will ensure all of our desired properties for the next marking. We can continue in this fashion indefinitely, thereby implementing our dynamical controller, provided that the system does not arrive at some deadlock situation created by 2 or more forbidden condition or initial state expressions.

Our task net could contain deadlock conditions created by the disabling of certain transitions for the previously mentioned expressions . In this case, the system is unrecoverable within the present framework. See the following figure.

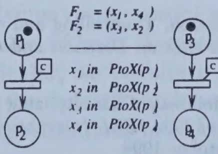


Fig. 6. Deadlock examples

In figure 6, if either controllable transition is fired, then the marking will not be forbidden condition free. The system is deadlocked since neither transition can fire. A similar condition exists for the $Expr_{init}$ expressions and the Task net, although we believe these will largely occur due to poor process design. An opportunity exists for handling such deadlock situations automatically. Since we generate the Task net off-line, we could perform some rigorous analysis of the system to determine areas of possible deadlock and modify our control expressions accordingly. This brings up some interesting issues for future research.

V. AREAS OF FUTURE RESEARCH AND CONCLUSION

This work is introductory and, in as much, the modeling power is limited. To improve this we would like to add controllable choice and synchronized event occurrences among other things to the task net framework. Similarly, we also want to generalize the types of subsystems we consider. In particular, certain subsystems can perform multiple tasks at one time(e.g. a conveyor which may move more than one part when actuated, or a carousel). These types of subsystems pose new difficulties such as: tasks must be simultaneously executed when state enabled; and preemption needs to be addressed. These are issues that are currently being addressed.

Another principle area of future research involves the synthesis of Task nets using the similar techniques to those outlined in [1]. Current research involves synthesizing sequential programs from condition-event models off line. These sequential programs could be our task sequences. A task net string could be generated by a modified algorithm similar to this work. There are other possibilities which we will not cover. Regardless, we hope that the task net structure is 'synthesize-able' and yet provides the dynamics required in the control of a system such as a automated manufacturing cell.

In summary, we have presented a possible abstraction for use in the control of manufacturing systems. This is based on a state based model and a set of resource objectives common to the types of subsystems we consider. We also implement a forbidden condition avoidance controller based along similar lines. We preserve the objectives that we desire but we do not yet ensure that the system will not deadlock.

REFERENCES

[1] L. E. Holloway and S. Rajaram. Synthesizing control code for achieving state targets in composed systems. In *Proceedings of 34th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, October 1996.

[2] P. J. Ramadge and W. M. Wonham. Supervisory control of a class of discrete-event processes. *SIAM J. on Control and Optimization*, 25(1):206–230, January 1987.

[3] L. E. Holloway and B. H. Krogh. Efficient synthesis of control logic for a class of discrete event systems. *1989 American Control Conference*, June 1989.

[4] Y. Li and W. M. Wonham. A state-variable approach to the modeling and control of discrete-event systems. *Proc. 26th Annual Allerton Conference*, pages 1140–1149, 1988. University of Illinois, Urbana.

[5] Ratnesh Kumar and Vijay K. Garg. *Modeling and Control of Logical Discrete Event Systems*. Kluwer Academic Publishers, 1995.

[6] Hao Zhong and W. Murray Wonham. On the consistency of hierarchical supervision in discrete-event systems. *IEEE Trans. on Automatic Control*, 35(10):1125–1134, 1990.

[7] M. Tittus, M. Fabian, and B. Lennartson. Controlling and coordinating recipes in batch applications. In *Proceedings of IEEE Conference on Decision and Control*, New Orleans, LA, December 1995.

[8] Jeffrey Ashley Jr. and Lawrence E. Holloway. Characterizing uncontrollable reachability for colored controlled petri nets. In *Proceedings of 1994 IEEE International Conference on Systems, Man, and Cybernetics*, San Antonio, Texas, October 1994.

[9] L.E. Holloway, X. Guan, and L. Zhang. A generalization of state avoidance policies for controlled Petri nets. *IEEE Trans. on Automatic Control*, page to appear, June 1996.

[10] T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, April 1989.

Multiple Marked Languages and Noninterference in Modular Supervisory Control*

J.G. Thistle and R.P. Malhamé
Département de génie électrique et de génie informatique
École Polytechnique de Montréal
C.P. 6079, Succ. Centre-ville
Montréal, (Québec)
Canada H3C 3A7

Abstract

Potential applications in distributed systems have motivated some useful extensions of basic supervisory control concepts. These extensions are outlined here. In particular, the original means of specifying minimal acceptable behaviour is replaced by a multiplicity of nonblocking conditions – this approach appears to be more conducive to modular specification; supervisor synthesis algorithms are then generalized accordingly; finally, a means of supervisor composition is proposed that generalizes the standard decentralized control schemes – this allows overall system functionality to increase rather than decrease under joint supervision.

1 Introduction

There is a growing trend toward the application of formal methods to service development in telecommunication networks and other distributed systems. This is illustrated by the so-called "feature interaction problem." Telecommunications call processing features and services (call waiting, multiway calling, etc.), developed and implemented in modular and decentralized fashion, sometimes interact in unforeseen and undesirable ways [1]. The practical consequences of such interference can be considerable, and the problem has recently generated significant research effort [5].

While feature interaction is a multi-facetted and as yet ill-defined problem, in some respects it appears to constitute an appropriate application for discrete-event control theory. Some preliminary work on its formalization within the framework of supervisory control has already been done (see for example [13, 3, 7]).

This initial confrontation of supervisory control with problems of service development has inspired some useful extensions of basic theoretical concepts [14]. These are outlined in the present report. In particular, we introduce a multiplicity of marked languages: these may constitute parts of the "plant" model or they may represent specifications Indeed, we shall only consider specifications in the form of nonblocking requirements, as

---

*Research supported by a grant from the Fonds pour la formation de chercheurs et l'aide à la recherche of Quebec, Canada, Bell-Northern Research Ltd. and the Natural Sciences and Engineering Research Council of Canada, under the Action concertée sur les méthodes mathématiques pour la synthèse de systèmes informatiques.

these can serve to specify both legal behaviour and a minimal degree of functionality. Examples of service development in telecommunications networks have shown that such multiple marked languages represent a particularly useful means of specifying minimal functionality within the context of modular control [13].

Secondly, we consider the synthesis of nonmodular, or "monolithic," supervisors that are nonblocking with respect to multiple marked languages; for the regular case we give a synthesis algorithm based on a fixpoint computation – in the special case of a single marked language, this fixpoint computation yields a simple presentation of the standard Ramadge-Wonham algorithm [9].

Finally, we turn to modular combinations of supervisors. Here we state conditions under which a conjunction [18] of two supervisors, each nonblocking with respect to a corresponding set of marked languages, is itself a nonblocking supervisor with respect to the union of the two sets of marked languages. This leads to a definition of *noninterfering* languages. We then propose a generalized notion of supervisor conjunction; this construction extends the standard decentralized control setup [8, 4], employing, instead of memoryless "natural projections," the "reporter maps" of hierarchical control [19]. This means of supervisor composition allows overall system functionality to increase rather than decrease under joint supervision.

Motivating examples for each of these extensions can be found in [13, 14], and proofs of results in [14].

2 Supervisory control preliminaries

As usual, let Σ^* represent the set of all finite strings of symbols belonging to the finite alphabet Σ, including the empty string of length zero, denoted here by 1. A *formal language* is simply a subset of Σ^*. If s and t denote strings over Σ, then st stands for their concatenation.

We write $s \leq t$ if $s \in \Sigma^*$ is a prefix, or initial substring, of $t \in \Sigma^*$, and $s < t$ if s is a proper prefix of t. The set of all prefixes of strings belonging to a formal language $K \subseteq \Sigma^*$ is denoted by $\overline{K}$; K is said to be *prefix-closed* if $K = \overline{K}$.

We model a DES as a prefix-closed formal language $L \subseteq \Sigma^*$ over a finite *event alphabet* Σ. A set of *control patterns* $\Gamma \subseteq 2^\Sigma$ represents the collection of subsets to which the set of possible events can be restricted through control action. A *supervisor* is a map

$$f : L \to \Gamma \text{ (pfn)}$$

taking event sequences to control patterns. The action of a supervisor $f : L \to \Gamma$ on the DES $L \subseteq \Sigma^*$ is modelled by the sublanguage $L^f \subseteq L \subseteq \Sigma^*$, defined recursively as follows [10]:

$$\begin{aligned} &1 \in L^f \\ &s\sigma \in L^f \iff s \in L^f,\ s\sigma \in L,\ \&\ \sigma \in f(s) \quad (\forall s \in \Sigma^*,\ \sigma \in \Sigma) \end{aligned}$$

In the first clause, 1 denotes the empty string. We say that the supervisor f is *complete* if $L^f \subseteq \text{dom}(f)$, where $\text{dom}(f)$ stands for the domain of f; this ensures that the above definition of L^f makes physical sense.

For a given language $K \subseteq \Sigma^*$, define the *active event set* after $k \in \overline{K}$,

$$\Sigma_K(k) := \{\sigma \in \Sigma \ :\ k\sigma \in \overline{K}\}$$

A sublanguage $S \subseteq L \subseteq \Sigma^*$ is *controllable* w.r.t. $L \subseteq \Sigma^*$ if

$$(\forall s \in \overline{S})(\exists \gamma \in \Gamma)[\Sigma_L(s) \cap \gamma = \Sigma_S(s)]$$

This controllability property serves to characterize all behaviours that can be achieved under supervision:

Proposition 2.1 (Ramadge-Wonham-Golaszewski [9, 6])
Let $S \subseteq L$ be nonempty. There exists a complete supervisor $f : L \longrightarrow \Gamma$ such that $L^f = S \subseteq L$ iff the language S is prefix-closed and controllable w.r.t. L.

□

3 Multiple marked languages

We shall consider specifications in the form of "nonblocking" conditions. A language $K \subseteq \Sigma^*$ is *nonblocking* w.r.t. $M \subseteq \Sigma^*$ if

$$\overline{K} \subseteq \overline{\overline{K} \cap M}$$

The nonblocking property w.r.t. a given language $M \subseteq \Sigma^*$ is preserved under arbitrary unions [2]; if the family $\Gamma \subseteq 2^{\Sigma}$ is closed under union, then controllability w.r.t. L is also preserved under arbitrary unions [9]. Moreover, the empty language is (vacuously) controllable w.r.t. any $L \subseteq \Sigma^*$ and nonblocking w.r.t. any $M \subseteq \Sigma^*$. It follows that the class of sublanguages of a given $L \subseteq \Sigma^*$ satisfying either or both of these properties contains a supremal element.

Examples of modular service development suggest the use of a collection of marked languages M_i, $i \in I$ [13]. Some of these may be introduced as part of the basic "plant" model, and others as a means of specifying desired behaviour of the controlled system. For example, in [14, 13], a small (three-subscriber) telephone network is modelled. This "plant" marks only strings in which all requests for calls (req_{ij}) are followed either by connections (con_{ij}) or by busy signals (no-con_{ij} or no-con_{ii}): the system is nonblocking with respect to this language only if all requests are eventually "answered," either by a connection or by a busy signal. But specifications also take the form of nonblocking conditions. A specification of a call waiting service is represented by figure 1.

The left- and rightmost states of figure 1 are those in which call waiting is in use – one call is in progress and a second call to subscriber i is on hold. (The events on-h_x and flash_i represent subscribers going "on-hook" (hanging up) and executing "flash-hooks", respectively.) The marked specification language is just the set of all strings that take the automaton to one of those states. Thus the system is nonblocking with respect to this language just in case it is always possible for the feature to be used at some future time. This nonblocking requirement ensures that the functionality of call waiting is always available, without requiring that any particular string be generated by the system, as would be the case if a minimal acceptable sublanguage were stipulated. This less rigid requirement appears more conducive to modular specification.

A supervisor $f :\longrightarrow \Gamma$ is said to be *nonblocking* with respect to marked languages M_i, $i \in I$ if $L^f = \overline{\overline{L^f} \cap M_i} = \overline{L^f \cap M_i}$, for all $i \in I$. This leads to the following immediate generalization of basic results of Ramadge and Wonham [9] and Golaszewski and Ramadge [6]:

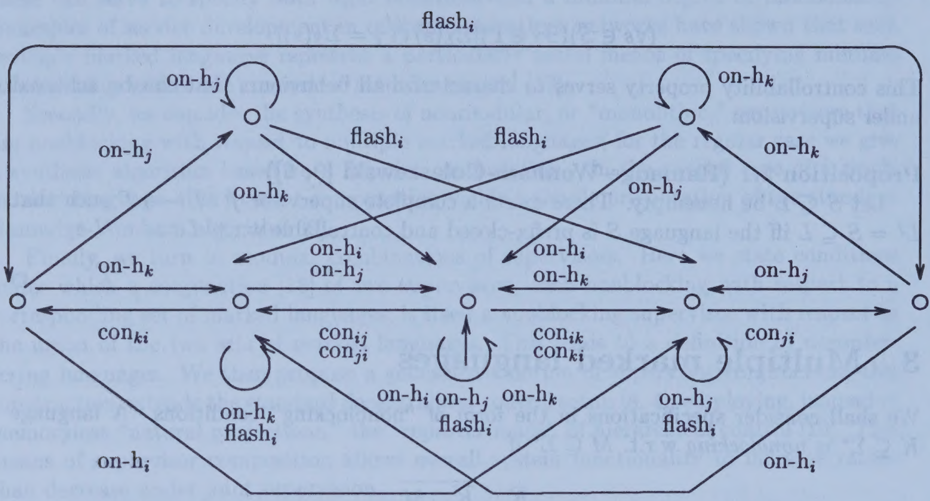

flash$_i$
on-h$_j$
on-h$_k$
on-h$_j$
flash$_i$
flash$_i$
on-h$_k$
on-h$_k$
on-h$_j$
on-h$_i$
on-h$_j$
on-h$_i$
on-h$_k$
on-h$_k$
on-h$_j$
con$_{ki}$
con$_{ij}$
con$_{ji}$
con$_{ik}$
con$_{ki}$
con$_{ji}$
on-h$_i$ on-h$_j$ on-h$_k$
flash$_i$
on-h$_k$
flash$_i$
on-h$_j$
flash$_i$
on-h$_i$
on-h$_i$

Figure 1: Call waiting specification [14, 13]. (Some transitions have been omitted for the sake of simplicity.)

Proposition 3.2 Let $K \subseteq L$ be prefix-closed and nonempty. Then there exists a complete, nonblocking supervisor $f : L \longrightarrow \Gamma$ such that $L^f = K$ iff

i. K is controllable with respect to L; and

ii. $K = \overline{K \cap M_i}$, for all $i \in I$.

□

Let $\mathcal{CN}(L; \{M_i : i \in I\})$ be the class of sublanguages of L that satisfy the two conditions of the above result. Like its counterpart in the standard framework, this class is nonempty and closed under arbitrary unions whenever $\Gamma \subseteq 2^{\Sigma}$ is closed under arbitrary unions. In this case it contains a supremal element, $\sup\mathcal{CN}(L; \{M_i : i \in I\})$. In the next section we show how to compute this supremal element in the case where all formal languages are regular. The computation yields a finite representation of a "minimally restrictive" supervisor f among those synthesizing languages L^f that are nonblocking w.r.t. M_i, $i \in I$.

4 Monolithic supervisor synthesis

In the regular language case, the problem of synthesizing a single, nonmodular supervisor reduces to that of controlling a finite automaton in such a way that it may always potentially enter any one of a collection of state subsets corresponding to the "marked" languages M_i, $i \in I$.

Formally, define a finite automaton to be (for our immediate purposes) a 6-tuple

$$\mathcal{A} = (\Sigma, Q, \delta, q_0, \Gamma, \{Q_i^m : i \in I\})$$

where

- Σ is a finite alphabet of event symbols;
- Q is a finite *state set*;
- $\delta : \Sigma \times Q \longrightarrow Q$ is a (partial) *transition function*;
- $q_0 \in Q$ is the *initial state*;
- $\Gamma \subseteq 2^{\Sigma}$ is a set of *control patterns*; and
- $\{Q_i^m \subseteq Q \ : \ i \in I\}$ is a set of *marked subsets*.

We assume without loss of generality that distinct state transitions carry distinct event symbols; that is,

$$(\forall \sigma \in \Sigma)(\forall q, q' \in Q)[\delta(\sigma, q)! \ \& \ \delta(\sigma, q')! \Longrightarrow q = q']$$

As usual, we extend the transition function δ to strings:

$$\delta : \Sigma^* \times Q \longrightarrow Q$$

We say that $\mathcal{A}$ *generates* a string $s \in \Sigma^*$ if $\delta(s, q_0)!$ (where the exclamation mark means that the preceding quantity is defined); we say that $\mathcal{A}$ *generates* s *under* a *feedback map* $f : \Sigma^* \longrightarrow \Gamma$ (pfn) if, for all $u < s$, $f(u)!$ and for all $\sigma \in \Sigma$,

$$u\sigma \leq s \Longrightarrow \sigma \in f(u)$$

The set of all strings generated by $\mathcal{A}$ under f is denoted by $L^f(\mathcal{A})$. The feedback map f is said to be *complete* with respect to $\mathcal{A}$ if $f(s)$ is defined for all $s \in \Sigma^*$ that is generated by $\mathcal{A}$ under f. (This ensures that the the definition of generation of strings under f makes sense.)

It will often be convenient to change the initial state of $\mathcal{A}$; let $\mathcal{A}_q$ be the automaton obtained by replacing q_0 with $q \in Q$.

Our computation of the supremal controllable and nonblocking sublanguage hinges on that of the set of initial states of a finite automaton $\mathcal{A}$ from which $\mathcal{A}$ can be controlled (by a feedback map) in such a way that it is always possible for it to enter any given Q_i^m, $i \in I$. Formally, define the state subset

$$\begin{aligned} F^{\mathcal{A}} := \{q \in Q \ : \ & (\exists f : \Sigma^* \longrightarrow \Gamma) \ f \text{ is complete for } \mathcal{A} \ \& \\ & (\forall s \in \Sigma^*)[s \in L^f(\mathcal{A}_q) \Rightarrow (\forall i \in I)(\exists t \in \Sigma^*)[st \in L^f(\mathcal{A}_q) \ \& \ \delta(st, q) \in Q_i]]\} \end{aligned}$$

In order to compute $F^{\mathcal{A}}$ it is convenient to introduce an operation that alters the automaton's control patterns [12]: for any $Q' \subseteq Q$, the automaton

$$\mathcal{A}(\not\to Q')$$

is obtained from $\mathcal{A}$ by replacing the set $\Gamma \subseteq 2^{\Sigma}$ of automaton control patterns with

$$\Gamma(\not\to Q') := \{\gamma' \subseteq \Sigma \ : \ (\exists \gamma \in \Gamma)[\gamma \setminus \{\sigma \in \Sigma : (\exists q \in Q)[\delta(\sigma, q) \in Q']\} \subseteq \gamma' \subseteq \gamma]\}$$

The effect of the operation is thus to allow feedback maps to disable any transitions leading to the state subset Q'.

We next adapt the *inverse dynamics operator* of [15, 12]:

$$\begin{aligned}\theta^{\mathcal{A}} :\ & 2^Q \longrightarrow 2^Q \\ & Q' \mapsto (\{q \in Q \ :\ (\exists \gamma \in \Gamma)(\forall \sigma \in \gamma)[\delta(\sigma, q)\ \not!]\} \cap Q') \\ & \qquad \cup \{q \in Q \ :\ (\exists \gamma \in \Gamma)[(\exists \sigma \in \gamma)[\delta(\sigma, q)!] \ \&\ (\forall \sigma \in \gamma)[\delta(\sigma, q)! \Rightarrow \delta(\sigma, q) \in Q']]]\}\end{aligned}$$

Intuitively, $\theta^{\mathcal{A}}(Q') \subseteq Q$ is the set of states from which the automaton may be "controlled" so as to remain in a state within $Q' \subseteq Q$ or to enter Q' in a single transition.

Using the "μ-calculus" fixpoint notation[1] we define the *controlled nonblocking subset* of $\mathcal{A}$ as

$$CN^{\mathcal{A}} := \nu X_0.\ \bigcap_{i \in I} \mu X_1.\ \theta^{\mathcal{A}(\not\to X_0)}(X_0 \cap (Q_i^m \cup X_1))$$

According to this notation, an expression prefixed with νX_0 denotes the greatest state subset $X \subseteq Q$ (in the sense of set inclusion) that is a "fixpoint" of that expression (in the sense that X equals the expression evaluated at $X_0 = X$); likewise, an expression preceded by the μX_1 "quantifier" denotes the least fixpoint. (Existence of such extremal fixpoints follows from the monotonicity of the expressions with respect to the quantified variables [11].)

The greatest fixpoint in the above expression is computed by iterating on X_0, starting with the initial value $X_0 = Q$; likewise, for a given value of X_0, the least fixpoint is computed by iteration on X_1, starting with the initial value $X_1 = \emptyset$ [11]. To understand the meaning of the expression for $CN^{\mathcal{A}}$, consider first the effect of the initial assignment $X_0 = Q$. All transitions of the automaton $\mathcal{A}(\not\to Q)$ can be prevented by a feedback map; it follows that

$$\begin{aligned} & \mu X_1.\ \theta^{\mathcal{A}(\not\to Q)}(Q \cap (Q_i^m \cup X_1)) \\ = \ & \mu X_1.\ \theta^{\mathcal{A}(\not\to Q)}(Q_i^m \cup X_1)\end{aligned}$$

is precisely the set of states from which a path (possibly of length zero) leads to Q_i^m. Indeed, to compute the least fixpoint, we first set $X_1 = \emptyset$, yielding

$$\theta^{\mathcal{A}(\not\to Q)}(Q_i^m)$$

or the set of states from which Q_i^m can be reached in a single transition; in general, the ith iteration yields the set of states from which the target subset Q_i^m can be reached in after at most i transitions. The least fixpoint is thus precisely the subset from which the target set can be reached in a finite number of steps.

It follows that

$$\bigcap_{i \in I} \mu X_1.\ \theta^{\mathcal{A}(\not\to Q)}(Q \cap (Q_i^m \cup X_1))$$

is the set of states from which $\mathcal{A}$ can be controlled so that any Q_i^m can eventually be reached. Iterating on the variable X_0 yields subsets of the form

$$\bigcap_{i \in I} \mu X_1.\ \theta^{\mathcal{A}(\not\to X_0)}(X_0 \cap (Q_i^m \cup X_1))$$

This expression denotes the set of states from which $\mathcal{A}$ can be controlled so that any subset $X_0 \cap Q_i^m$ can eventually be reached, by way of a path that lies within the subset X_0.

---

[1] See [15] for the relevant bibliography.

Fixpoint theory [11] states that the greatest fixpoint

$$CN^{\mathcal{A}} = \nu X_0.\ \bigcap_{i\in I} \mu X_1.\ \theta^{\mathcal{A}(\not\to X_0)}(X_0 \cap (Q_i^m \cup X_1))$$

is also the greatest "post-fixpoint"

$$X_0 \subseteq \bigcap_{i\in I} \mu X_1.\ \theta^{\mathcal{A}(\not\to X_0)}(X_0 \cap (Q_i^m \cup X_1))$$

This is simply the largest subset $X_0 \subseteq Q$ from which the automaton can be controlled in such a way that there always exists a path through X_0 to any $X_0 \cap Q_i^m$; by extension, this subset is thus the largest from which $\mathcal{A}$ can be controlled so that all of the Q_i^m remain reachable:

Proposition 4.3
Let $\mathcal{A}$ be an automaton of the form defined above. Then

$$CN^{\mathcal{A}} = F^{\mathcal{A}}$$

□

The above fixpoint characterization allows us to compute $F^{\mathcal{A}}$ by iteration of the appropriate operators – this calculation requires time polynomial in the size of the state set of $\mathcal{A}$ and linear in the number of subsets Q_i^m. The computation yields as a byproduct a feedback map that applies suitable control from any initial state in $F^{\mathcal{A}}$ (see the proof of the above proposition in the appendix). If $\Gamma \subseteq 2^{\Sigma}$ is closed under union, then once $F^{\mathcal{A}}$ is computed we have an automaton representation of $\sup \mathcal{CN}(L)$:

Proposition 4.4
Let $\mathcal{A} = (\Sigma, Q, \delta, q_0, \Gamma, \{Q_i^m \subseteq Q\ :\ i \in I\})$ be a finite automaton as defined above with $\Gamma \subseteq 2^{\Sigma}$ closed under arbitrary unions, and suppose that $L = \{s \in \Sigma^*\ :\ \delta(s, q_0)!\}$ and $M_i = \{s \in \Sigma^*\ :\ \delta(s, q_0) \in Q_i^m\}$ for all $i \in I$. If $q_0 \notin CN^{\mathcal{A}}$, then $\sup \mathcal{CN}(L, \{M_i\ :\ i \in I\}) = \emptyset$. Otherwise, $\sup \mathcal{CN}(L, \{M_i\ :\ i \in I\})$ is just the language accepted by the finite automaton

$$\mathcal{A}_{CN} = (\Sigma, CN^{\mathcal{A}}, q_0, \delta_{CN}, CN^{\mathcal{A}})\,,$$

where

$$\delta_{CN}:\ \Sigma \times CN^{\mathcal{A}} \longrightarrow CN^{\mathcal{A}}$$

$$(\sigma, q) \mapsto \begin{cases} \delta(\sigma, q) & \text{if } \sigma \in \bigcup\{\gamma' \in \Gamma\ :\ (\forall \sigma' \in \gamma')[\delta(\sigma', q) \in CN^{\mathcal{A}}]\} \\ \text{undefined} & \text{otherwise} \end{cases}$$

(That is, $\sup \mathcal{CN}(L) = \{s \in \Sigma^*\ :\ \delta_{CN}(s, q_0) \in CN^{\mathcal{A}}\} = \{s \in \Sigma^*\ :\ \delta_{CN}(s, q_0)!\}$.)

□

Indeed, if Γ is closed under union, then $\mathcal{A}_{CN}$ can be taken as a representation of a complete supervisor that synthesizes $\sup\mathcal{CN}(L; \{M_i\ :\ i \in I\})$.

While supervisor synthesis can be carried out in time polynomial in the size of the state set of the automaton $\mathcal{A}$, if the marked languages M_i are initially specified by a number of distinct component automata, then the cardinality of the state set of $\mathcal{A}$ of course increases exponentially with the number of such components. This is one of many reasons for considering modular supervision [18].

5 Modular synthesis

In this section we present the appropriate generalizations of the basic results on modular supervisor synthesis [18].

Let f_1 and f_2 be supervisors, and let the set of control patters $\Gamma \subseteq 2^{\Sigma}$ be closed under intersection. The *conjunction* [18] of f_1 and f_2,

$$\begin{aligned} f_1 \wedge f_2 \ &\longrightarrow \Gamma \qquad\qquad \text{(pfn)} \\ s \ &\mapsto f_1(s) \cap f_2(s) \end{aligned}$$

(which maps each string to the intersection of the control patterns selected by f_1 and f_2), is itself a supervisor. This standard means of composing supervisors models the effect of running the two "in parallel," and according each of them "veto power" over event occurrences. Naturally, even if the individual supervisors are nonblocking with respect to some given marked languages, their conjunction need not be:

Proposition 5.5 If $\Gamma \subseteq 2^{\Sigma}$ is closed under intersection and $f_1 : L \longrightarrow \Gamma$ and $f_2 : L \longrightarrow \Gamma$ are nonblocking with respect to the sets of marked languages $\{M_i : i \in I\}$ and $\{M_j : j \in J\}$, respectively, then $f_1 \wedge f_2$ is a nonblocking supervisor with respect to the marked languages $\{M_k : k \in I \cup J\}$ if and only if

$$L^{f_1} \cap L^{f_2} = \overline{L^{f_1} \cap L^{f_2} \cap M_k}\ , \ \forall k \in I \cup J\ ,$$

□

That is, if the set of control patterns is closed under intersection, and f_1 and f_2 are nonblocking with respect to their respective sets of marked languages, then $f_1 \wedge f_2$ is a nonblocking supervisor with respect to all of the marked languages if and only if $L^{f_1} \cap M_i$ and L^{f_2} are nonconflicting for all $i \in I$ and L^{f_1} and $L^{f_2} \cap M_j$ are nonconflicting for all $j \in J$. – under these conditions, we say that the languages L^{f_1} and L^{f_2} are *noninterfering* with respect to $\{M_k : k \in I \cup J\}$. Intuitively, the prefix-closed languages synthesized by two supervisors are noninterfering if neither supervisor prevents the other from achieving any of its control objectives.

The automaton of figure 2 represents a specification of a three-way calling feature for the small telephone network of [14, 13]. The marked language is the set of all strings taking the automaton to its topmost state, which means that a three-way call is in progress. The respective minimally restrictive supervisors for the call waiting and three-way calling specifications interfere with each other; indeed, after a first call is established, the call waiting supervisor permits only incoming calls, and the three-way calling supervisor allows only outgoing calls. This is a simple version of a well-known example of feature interaction; more examples are given in [14, 13]. Indeed, the notion of interfering supervisors appears to characterize many feature interactions.

The following result shows that, under the noninterfering condition, supervisor conjunction preserves "minimal restrictiveness" of supervisors (cf. [18]):

Proposition 5.6 Let $\Gamma \subseteq 2^{\Sigma}$ be closed under intersection and let $f_1 : L \longrightarrow \Gamma$ and $f_2 : L \longrightarrow \Gamma$ be nonblocking supervisors such that $L^{f_1} = \sup\mathcal{CN}(L, \{M_i : i \in I\})$ and $L^{f_2} = \sup\mathcal{CN}(L, \{M_j : j \in J\})$. Then $L^{f_1 \wedge f_2} = \sup\mathcal{CN}(L, \{M_k : k \in I \cup J\})$ if and only if L^{f_1} and L^{f_2} are noninterfering with respect to $\{M_k : k \in I \cup J\}$.

□

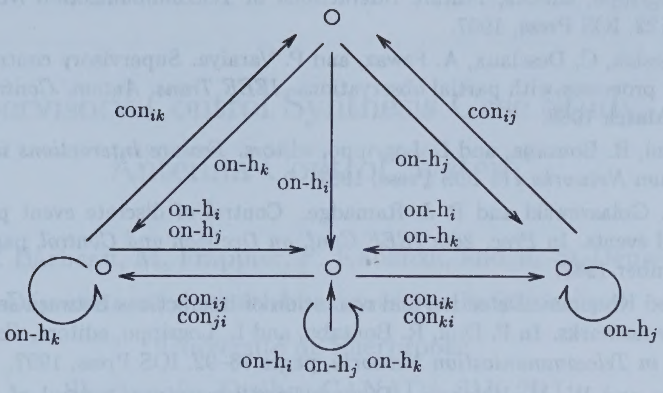


Figure 2: Three-way calling specification [14, 13]. (Some transitions have been omitted for the sake of simplicity.)

Conclusion

The use of multiple marked languages can be a natural means of specifying supervisory control problems. We have illustrated this point and extended some basic results on monolithic and modular synthesis.

The call waiting / three-way calling example illustrates that ordinary conjunction is in some instances an inappropriate means of composing modular supervisors: a subscriber to call waiting and multiway calling should enjoy freer behaviour under the two services than under a single one. In [17, 16], reporter maps such as those used to model hierarchical control are interposed between respective supervisors and the plant before the supervisors are conjoined, filtering their observations and control actions. This more general means of supervisor composition can result in the plant having greater liberty under joint supervision than under the direct control of a single supervisor. Indeed, the construction of reporter maps having suitable "observer" properties is a means of resolving interference among supervisors [17, 16]. This potential paradigm for the resolution of feature interactions is a current topic of research.

References

[1] E. Jane Cameron, Nancy Griffeth, Yow-Jian Lin, Margaret E. Nilson, and William K. Schnure. A feature interaction benchmark in IN and beyond. *IEEE Communications Magazine*, 31(3), 1993.

[2] Enke Chen and Stéphane Lafortune. On nonconflicting languages that arise in supervisory control of discrete event systems. *Systems and Control Letters*, 17:105–113, 1991.

[3] Yi-Liang Chen, Stéphane Lafortune, and Feng Lin. Resolving feature interactions using modular supervisory control with priorities. In P. Dini, R. Boutaba, and

L. Logrippo, editors, *Feature Interactions in Telecommunication Networks*, pages 108–122. IOS Press, 1997.

[4] R. Cieslak, C. Desclaux, A. Fawaz, and P. Varaiya. Supervisory control of discrete-event processes with partial observations. *IEEE Trans. Autom. Control*, 33(3):249–260, March 1988.

[5] P. Dini, R. Boutaba, and L. Logrippo, editors. *Feature Interactions in Telecommunication Networks IV.* IOS Press, 1997.

[6] C. H. Golaszewski and P. J. Ramadge. Control of discrete event processes with forced events. In *Proc. 26th IEEE Conf. on Decision and Control*, pages 247 – 251, December 1987.

[7] Ahmed Khoumsi. Detection and resolution of interactions between services of telephone networks. In P. Dini, R. Boutaba, and L. Logrippo, editors, *Feature Interactions in Telecommunication Networks*, pages 78–92. IOS Press, 1997.

[8] F. Lin and W. M. Wonham. Decentralized supervisory control of discrete-event systems. *Information Sciences*, 44:199–224, 1988.

[9] P. J. Ramadge and W. M. Wonham. Supervisory control of a class of discrete event processes. *SIAM J. Control and Optimization*, 25(1):206–230, 1987.

[10] Peter J. G. Ramadge. Some tractable supervisory control problems for discrete-event systems modeled by Büchi automata. *IEEE Trans. Automatic Control*, 34(1):10–19, January 1989.

[11] Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285–309, 1955.

[12] J. G. Thistle. On control of systems modelled as deterministic Rabin automata. *Discrete Event Dynamic Systems: Theory and Applications*, 5(4):357–381, September 1995.

[13] J. G. Thistle, R. P. Malhamé, H.-H. Hoang, and S. L. Lafortune. Feature interaction modelling, detection and resolution: A supervisory control approach. In P. Dini, R. Boutaba, and L. Logrippo, editors, *Feature Interactions in Telecommunication Networks*, pages 93–107. IOS Press, 1997.

[14] J. G. Thistle, R. P. Malhamé, H.-H. Hoang, and S. L. Lafortune. Supervisory control of distributed systems, Part I: Modelling, specification and synthesis. Technical Report EPM/RT-97/08, École Polytechnique de Montréal, February 1997.

[15] J. G. Thistle and W. M. Wonham. Supervision of infinite behaviour of discrete-event systems. *SIAM J. Control and Optimization*, 32(4):1098–1113, July 1994.

[16] K. C. Wong, J. G. Thistle, H.-H. Hoang, and R. P. Malhamé. Conflict resolution in modular control with applications to feature interaction. In *Proc. 34th IEEE Conference on Decision and Control*, pages 416–421, December 1995.

[17] K. C. Wong, J. G. Thistle, and R. P. Malhamé. Conflict resolution with flexible priority in modular control. In *Canadian Conference on Electrical and Computer Engineering*, pages 797–800, September 1995.

[18] W. M. Wonham and P. J. Ramadge. Modular supervisory control of discrete-event systems. *Mathematics of Control, Signals and Systems*, 1:13–30, 1988.

[19] Zhong Hao and W. Murray Wonham. On the consistency of hierarchical supervision in discrete-event systems. *IEEE Trans. Automatic Control*, 35(10):1125–1134, October 1990.

A Supervisory Control Synthesis Case Study: The Antenna Control System*

M. Barbeau, M. Frappier, F. Kabanza, and R. St-Denis
Département de mathématiques et d'informatique
Université de Sherbrooke
Sherbrooke, Québec CANADA J1K 2R1

Abstract

We devised an algorithm for deriving controllers for timed discrete-event systems with nonterminating behavior modeled by timed transition graphs and control requirements expressed by Metric Temporal Logic (MTL) formulas. This algorithm has several interesting features. Firstly, it simultaneously handles the issues of controllability, safety, liveness, and real time in a single framework. Secondly, time and space complexity of the algorithm is reduced thanks to ingenious search and representation techniques. This algorithm has a general character and is implemented in a system called Temporal Controller Synthesis Tool (TCST).

In our research, there is a strong commitment to the application of theoretical research results. We validated our algorithm through case studies. This is very important for assessing its realism. This paper presents the application of our algorithm to the derivation of a controller for an Antenna Rotor Control System (ARCS). The ARCS is responsible for orienting antennas in the direction of a telecommunications satellite. The paper presents modeling of the rotor system with timed transition graphs, specification of the constraints with MTL, and a controller derived using TCST.

1 Introduction

Synthesis of controllers is defined as a process aiming at systematically deriving a model of the behavior of a supervisor given a specification of control requirements and a model

---

[*] The research described in this paper was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds pour la formation de chercheurs et l'aide à la recherche (FCAR).

of the perception of the behavior of a process. Assessment of the suitability of synthesis approaches on real control problems is important in order to uncover flaws and guide future research activities. This paper presents the development of a synthesis case study using an algorithm handling liveness, safety, and real-time control constraints. The control problem consists of deriving the kernel of an Antenna Rotor Control System (ARCS).

The ARCS is responsible for maintaining the orientation of antennas in the direction of a moving telecommunications satellite. The direction is expressed in terms of an azimuth and an elevation (both in degrees). Its central component is an Antenna Direction Controller (ADC) responsible of deciding when to start/stop rotors and, whenever they are running, the direction of their movement. For this purpose, the ADC is connected to specialized movement sensing and control processes. There are separate sensing and control processes for azimuth and elevation. The sensing elements inform about the position of the antennas relative to an azimuth target and an elevation target. The rotor controllers are responsible for the timing details involved in switching on and off motors of the rotor in one direction or the other. Finally, the ADC is also responsible for synchronizing the operation of the controllers with the operation of a larger satellite tracking process (the antennas are pictured in Figure 1). Rotors appear at the junction of the vertical and horizontal masts.

The architecture of ARCS is illustrated in Figure 2 (this design is inspired from a design of a traffic light control system that appears in [5]). A solution for the ADC is developed using a synthesis algorithm described in References [1] and [2]. This algorithm allows synthesis from control requirements that include safety, liveness, and real-time properties. The control requirements are expressed by Metric Temporal Logic (MTL) formulas and the nonterminating behavior of the system is modeled by a timed transition graph. Events have time duration and states are labeled by propositional symbols. This algorithm is also quite close to one described in [3], except that the latter is formulated in an artificial intelligence planning paradigm.

The rest of this paper is structured as follows. Section 2 presents the model of the perception of the process and the control requirements. Section 3 discusses a solution. We conclude with Section 4.

2 Modeling of ADC

We model a process as a timed transition graph $G = (X, \mathcal{P}, \lambda, A, \tau, \xi, x_0)$, where X is a finite set of states; $\mathcal{P}$ is a finite set of propositional symbols; $\lambda : X \to 2^{\mathcal{P}}$ is a labeling function that assigns to each state the set of propositional symbols true at that state; A is a finite set of actions; $\tau : A \to \mathbb{R}^+$ is a time duration function such that $\tau(a) > 0$

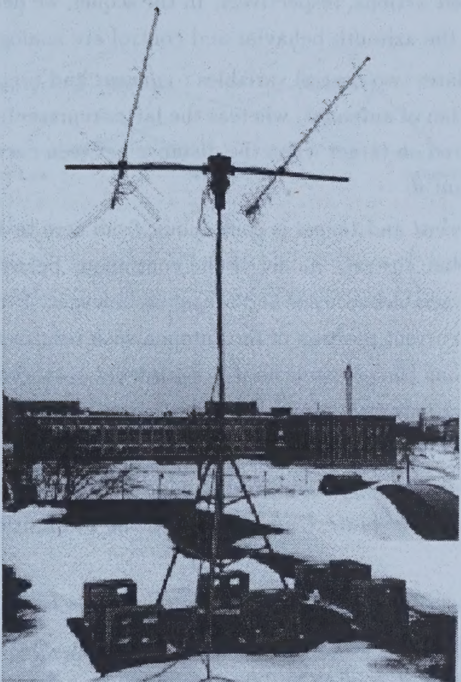


Figure 1: Antennas of a satellite tracking system

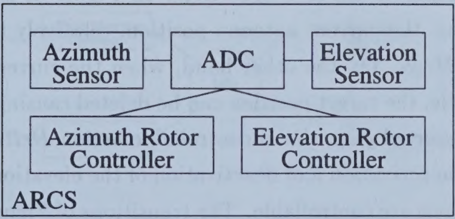

Azimuth Sensor
ADC
Elevation Sensor
Azimuth Rotor Controller
Elevation Rotor Controller
ARCS

Figure 2: Architecture

for all $a \in A$; $\xi : X \times A \to X$ is a transition function; and $x_0 \in X$ is an initial state. The set A is partitioned into the sets A_c and A_{uc} denoting the set of *controllable* actions and set of *uncontrollable* actions, respectively. In the sequel, we detail only the elevation behavior and control, the azimuth behavior and control are analogous.

The ADC manipulates two control variables : *current* and *target*. The former represents the current position of antennas, whereas the latter represents its targeted position. An antenna is considered on target when the distance between *current* and *target* is less or equal than a constant d.

The domain of *current* and *target* is continuous, from zero to 180 degrees. However, we only reason about an abstract model of the continuous behavior in which only the relations between the variables *current* and *target* are relevant. Propositions *Low*, *Good*, and *High* refer to the current position of the antenna with respect to the target and they hold when the conditions $(target - current > d)$, $|target - current| \leq d$, or $(current - target > d)$ are respectively true. Initially, the relation between *target* and *current* is unknown which is conveyed by the proposition *Unknown*.

The propositions *Idle*, *Moving Down*, and *Moving Up* refer to the state of the elevation rotor. The value of $\lambda(x_0)$ is $\{Idle, Unknown\}$, i.e., the propositions true in the initial state x_0.

For the ADC, the set of propositional symbols $\mathcal{P}$ contains *Low*, *Good*, *High*, *Unknown*, *Idle*, *Moving Down*, and *Moving Up*.

The transition function ξ is often better represented by a diagram. We use Petri nets diagrams in which circles represent propositions and rectangles represent transitions. Transitions are connected to propositions. Incoming propositions are retracted, outgoing propositions are asserted, and propositions connected with arrows at both sides are tested by the transition. In our model, every action has a duration of one time unit.

The diagrams for the ADC are quite simply described. Target specification can be performed solely when propositions *Idle* and *Unknown* are true (i.e., the uncontrollable actions *Setpos Low*, *Setpos Good*, or *Setpos High*, on Figure 3, is executed). This has three possible outcomes: the current antenna position, relatively to the target, is set to either *Low*, *Good*, or *High*. On the other hand, when the current antenna position is good and the rotor is idle, the target position can be deleted causing *Good* to be retracted and *Unknown* to be asserted (i.e., the uncontrollable action *Deltpos* is executed). The transitions modeling the activation and deactivation of the elevation rotor are pictured in Figure 4. All these actions are controllable. The transitions of Figure 5 model the current antenna position monitoring, while the rotor is active. All these actions are controllable except action *Wait* which models inactivity.

The control requirements are specified with Metric Temporal Logic (MTL) [4], in which time constraints are associated with modal operators. This allows the specification

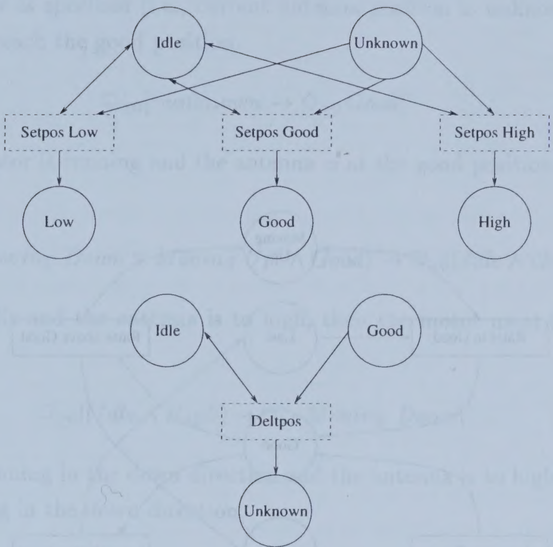

Idle
Unknown
Setpos Low
Setpos Good
Setpos High
Low
Good
High
Idle
Good
Deltpos
Unknown

Figure 3: Setting and deleting target position

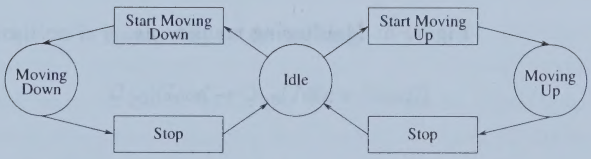

Start Moving Down
Start Moving Up
Moving Down
Idle
Moving Up
Stop
Stop

Figure 4: Rotor activation and deactivation transitions

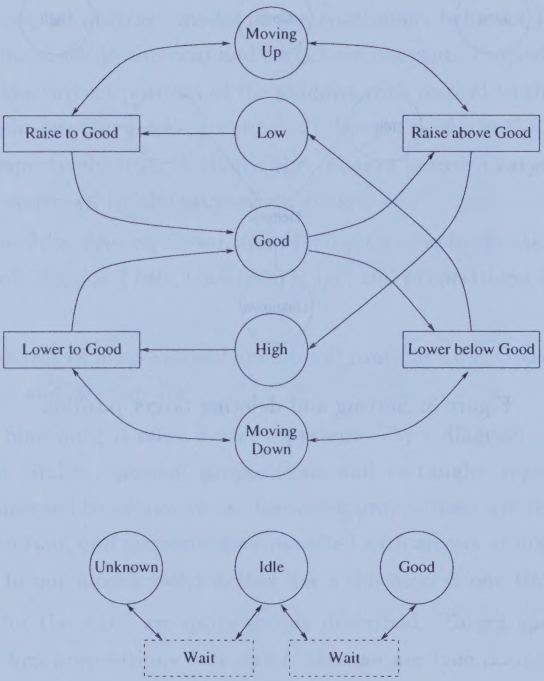

Moving Up
Raise to Good
Low
Raise above Good
Good
Lower to Good
High
Lower below Good
Moving Down
Unknown
Idle
Good
Wait
Wait

Figure 5: Monitoring transitions

of various properties such as "eventually, within t time units, property p will be satisfied" or "property p must always be satisfied after t time units." The following formulas specify the elevation constraints.

When a target is specified (i.e., current antenna position is unknown), the antenna must eventually reach the good position.

$$\Box_{\geq 0}[\neg unknown \rightarrow \Diamond_{\geq 0} Good] \tag{C1}$$

Whenever the motor is running and the antenna is at the good position, the motor must be stopped.

$$\Box_{\geq 0}[((Moving\ Down \vee Moving\ Up) \wedge Good) \rightarrow \bigcirc_{\geq 0}(Idle \wedge Good)] \tag{C2}$$

If the motor is idle and the antenna is to high, then the motor must be started in the down direction.

$$\Box_{\geq 0}[(Idle \wedge High) \rightarrow \bigcirc_{\geq 0} Moving\ Down] \tag{C3}$$

If the motor is running in the down direction and the antenna is to high, then the motor must keep running in the down direction.

$$\Box_{\geq 0}[(Moving\ Down \wedge High) \rightarrow \bigcirc_{\geq 0} Moving\ Down] \tag{C4}$$

If the motor is idle and the antenna is to low, then the motor must be started immediately in the up direction.

$$\Box_{\geq 0}[(Idle \wedge Low) \rightarrow \bigcirc_{\geq 0} Moving\ Up] \tag{C5}$$

If the motor is running in the up direction and the antenna is to low, then the motor must keep running in the up direction.

$$\Box_{\geq 0}[(Moving\ Up \wedge Low) \rightarrow \bigcirc_{\geq 0} Moving\ Up] \tag{C6}$$

If the antenna position is good, wait or do nothing.

$$\Box_{\geq 0}[Good \rightarrow \bigcirc_{\geq 0}(Idle \vee Good)] \tag{C7}$$

3 Synthesis of a Solution

We discuss synthesis of the controller ADC using an algorithm presented in References [1] and [2]. This algorithm produces a controller by using a forward search technique, combined with incremental model checking. More specifically, this is a combination of three main operations:

- Incremental exploration of the global state space of the interleaved processes, while verifying the MTL formulas over trajectories (sequences of states) to detect bad states (violations of safety properties) or bad cycles (violations of liveness properties). The result is a graph of global states labeled by MTL formulas.

- Use of a control-directed backtracking technique that goes back over uncontrollable paths of arbitrary but finite length, from bad states or states that close bad cycles, to prune the search space more efficiently. Most of the states on these paths are not expanded further.[1]

- Incremental creation of the state space of the controller and calculation of the feedback function by determining incrementally states at which controllable events must be enabled (this is dual to determining those that must be disabled).

The obtained controller is represented by a pair $(\mathcal{M}, \phi)$, where $\mathcal{M} = (Q, A, \delta, q_0)$ is a transition structure and $\phi : Q \to \Gamma$ a feedback function determining enabled actions. The combination of a process and a controller constitutes a closed-loop system. In fact, $\mathcal{M}$ mimics the behavior of the concurrent execution of processes while ϕ determines the set of permissible process actions for each step of the execution of the closed-loop system. The reader is referred to [1] and [2] for more details. Computation has been performed with our own software tool, called Temporal Controller Synthesis Tool (TCST), implemented in Lisp.

Figure 6 illustrates the transition structure synthesized for ADC. States are represented as circles embedding propositional symbols. Initially, the controller is in state 1. From this state, it senses the current antenna position and compares the result with the target. The outcome is uncontrollable and the controller may handle any of the three possible alternatives (high, good, or low).

Let us consider in more detail the case where the position is high. In state 2, the precondition of constraint C1 is satisfied. The transition structure of Figure 6 presents two choices of trajectories that may be selected according to the fact that either constraints C3 is taken into account or not by the synthesis algorithm. If C3 is taken into account, the precondition of C3 is true in state 2. Satisfaction postcondition of C3 is required in the next state. In other words, activation of the rotor in the down direction is forced. The controller moves to state 6. The antenna is eventually lowered to the good position, moving from state 6 to state 7. Because of constraint C2, the rotor is stopped, moving from state 7 to state 3.

The process remains in state 3 (action *Wait*) until a new sensing of the current antenna position is required (action *Deltpos*) which takes the controller back to the initial state.

---

[1] The control-directed backtracking technique is the main difference between a similar algorithm described in [3], where uncontrollable transitions are abstracted over by nondeterministic transitions.

Now, let us consider a scenario in which constraint C3 is not interpreted by the synthesis algorithm. It is perfectly valid, although not optimal, for the controller to behave as follows from state 2. Start the rotor in the up direction for a while (transition from state 2 to state 5). Stop the rotor (transition from state 5 to state 10). Then, finally, start the rotor in the down direction (transition from state 10 to state 6).

The point to stress is here is that in the MTL formula modeling the control requirements, there may be some sub formulas that may not be required to insure synthesis of a valid solution. They are, however, helpful because they represent heuristics guiding the algorithm to synthesize more quickly a less complex solution. Indeed, in the current example, solely constraints C1 is mandatory for the validity of a solution. Although, conjunction of C1 with C2 to C7 contributes to limit the number of generated states and compute a more optimal solution.

In Figure 6, dotted arrows represent additional trajectories generated if only C1 is considered (the dashed arrow must not be considered in this solution). A total of ten states is generated. Conjunction with C2 to C7 leads to a controller without the dotted transitions and states 5 and 10 (but with the dashed transition). Hence, the controller has only 8 states.

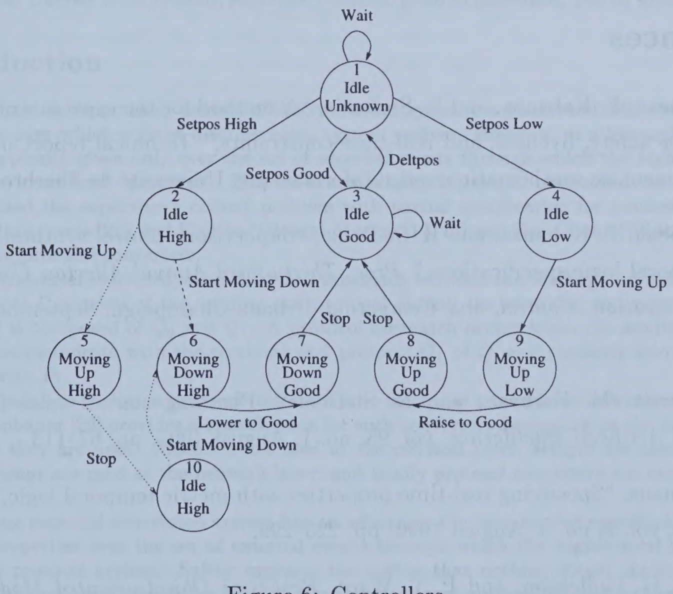

Wait
1
Idle
Unknown
Setpos High
Setpos Low
Deltpos
Setpos Good
2
Idle
High
3
Idle
Good
Wait
4
Idle
Low
Start Moving Up
Start Moving Down
Start Moving Up
Stop
Stop
5
Moving
Up
High
6
Moving
Down
High
7
Moving
Down
Good
8
Moving
Up
Good
9
Moving
Up
Low
Lower to Good
Raise to Good
Start Moving Down
Stop
10
Idle
High

Figure 6: Controllers

The feedback function ϕ can be easily computed from the transition structure of the controller. Actions that are permissible in a given state are those labeling transition outgoing from that state. Note that if, in a given state, there are active uncontrollable actions, only them are permitted in that state. Otherwise, only one controllable action is permitted. In this experiment, TCST did not compute a maximal controller in the traditional control theoretical sense (i.e., permit as many actions as possible). The controllable actions may be optimal though in a practical sense: e.g., enabling actions with the lowest cost for instance. This is in the line of AI or decision optimization concerns, as discussed in [3]. Nevertheless, it is possible to run TCST with an optional parameter requiring it to compute a maximal solution whenever there exists one. As in [6], such a maximal solution does not always exists when liveness constraints are involved.

4 Conclusion

We presented a simple realistic synthesis example. Calculation of the controller has been performed with the TCST tool within a period of 0.13 seconds. This suggests that the synthesis approach is feasible. In addition, formalization of our problem before implementation helped us to uncover errors in our design.

Future work is required to improve the realism of our model. For example faults are not taken into account.

References

[1] M. Barbeau, F. Kabanza, and R. St-Denis, "A method for the synthesis of controllers to handle safety, liveness, and real-time constraints," Technical report number 196, Département de mathématiques et d'informatique, Université de Sherbrooke, 1997.

[2] M. Barbeau, F. Kabanza, and R. St-Denis, "Supervisory control synthesis from metric temporal logic specifications," *Proc. Thirty-third Annual Allerton Conference on Communication, Control, and Computing*, Urbana-Champaign, September 1995, pp. 96–105.

[3] F. Kabanza, M. Barbeau, and R. St-Denis, "Planning control rules for reactive agents," *Artificial Intelligence*, vol. 95, no. 1, August 1997, pp. 67–113.

[4] R. Koymans, "Specifying real-time properties with metric temporal logic," *Real-time Systems*, vol. 2, no. 4, August 1990, pp. 225–299.

[5] B. Selic, G. Gullekson, and P. T. Ward, *Real-time Object-oriented Modeling*, John Wiley & Sons, Inc., 1994.

[6] J. G. Thistle and W. M. Wonham, "Control of infinite behavior of finite automata," *SIAM J. Control and Optimization*, vol. 32, no. 4, pp. 1098–1113, 1994.

On Supervisory Control of Partial Specification arising in Protocol Conversion *

Ratnesh Kumar
Dept. of Elec. Eng., Univ. of Kentucky
Lexington, KY, and
Applied Research Lab., Penn State
State College, PA

Martin Fabian
Control Eng. Lab.
Chalmers Univ. of Tech.
Gothenberg, Sweden

Abstract

The protocol conversion problem requires synthesis of a supervisor so that the overall system *implements* a service specification defined over only the external set of events—a partial specification. In this paper we show that by suitably defining a family of marked languages, this problem can be reduced to the nonblocking supervisory control problem over the entire event set, so that the existing results and algorithms can be directly applied. We also provide techniques for minimally altering the service specification when the given specification cannot be implemented. Finally, we present an on-line control computation algorithm of linear complexity that implements the service specification effectively by executing only a minimal controlled plant behavior.

Keywords: Discrete event systems, supervisory control, protocol conversion, partial specification

1 Introduction

Most research on supervisory control of discrete event systems is concerned with specifications on the entire event set over which a given discrete event system evolves. However, in a hierarchical setting, specifications are typically given only over the set of *external* events through which the higher level "user" interacts with the lower level "provider". This for example holds in layered communication protocols [29, 27]. Overkamp studied the supervisory control problem with partial specification for nondeterministic systems modeled by failure traces [22], and Kumar-Nelvagal-Marcus [17, 18] and Hsu-Chang [10] studied this problem for the design of protocol converters.

A protocol converter is needed when there is a mismatch between the sending and the receiving ends. In Figure 1 protocol P consists of the sending end protocol P_0 and the receiving end protocol P_1. Similarly, the protocol Q is composed of Q_0 and Q_1. A protocol mismatch occurs when the sending end protocol P_0 of P needs to communicate with the receiving end protocol Q_1 of Q, and similarly also when Q_0 needs to communicate with P_1.

A practical solution to such a protocol mismatch is to interpose a translator or a converter as depicted in Figure 2. Tanenbaum [29] provides a classification for such converters depending on the layer of the protocol stack at which they are used. *Repeaters* are used at the physical layer; *bridges* are used at the link layer; *routers* or *gateways* are used at the network layer; and finally *protocol converters* are used at the transport layer.

The resulting protocol conversion system has to adhere to the user service specification given as *safety* and *progress* properties over the set of external events through which the higher level users interact with the lower level protocol system. Safety captures the notion that nothing illegal should happen, i.e., the event traces of the protocol conversion system should correspond to those allowed by the user specification. Progress captures the notion that the composition should not block the occurrence of an external event whenever it is feasible in the user specification, i.e., whenever the user specification does not block it.

Derivation of protocol converters has received considerable attention [5, 19, 1, 3, 14, 21, 23]. The protocol conversion problem can be conveniently formulated in the supervisory control framework which naturally

---

*This research was supported in part by the Center for Robotics and Manufacturing Systems, University of Kentucky, in part by the National Science Foundation under the Grants NSF-ECS-9409712 and NSF-ECS-9709796, in part by the Office of Naval Research under the Grant ONR-N00014-96-1-5026, and in part by the Swedish National Board for Industrial and Technical Development (NUTEK) under the Grant 9304792-5.

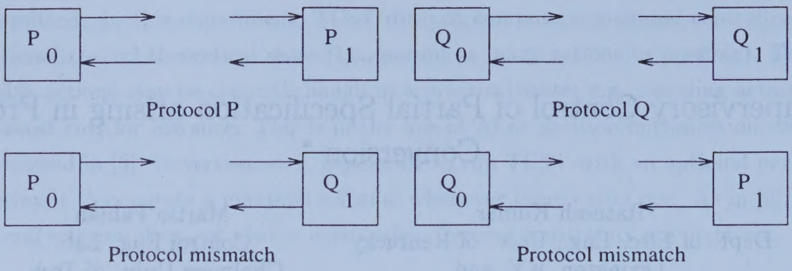


Figure 1: Protocol configuration

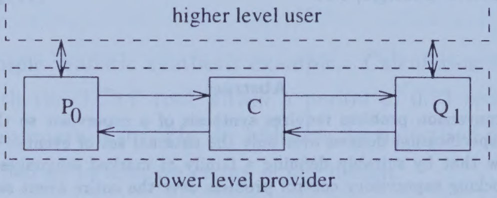


Figure 2: Interposing a protocol converter

captures the limited control and observation capabilities of a converter by means of uncontrollability and unobservability of events. Such a formulation was considered by Inan as an important application example in [11], in further depth in [12], and the theoretical foundation on which this application is based was given in [13]. However, these work of Inan *only* address the safety constraint of the protocol conversion problem which requires that the projected language of the supervised system should equal the given specification language. The additional progress constraint, which requires that the supervised system never block an external event that is not blocked by the specification itself, was *also* considered as part of the protocol conversion formulated in the work of Calvert-Lam [3]. In [17, 18], the problem of protocol conversion was studied as a discrete event supervisory control problem in which both the safety and progress constraints were considered. A necessary and sufficient condition for the existence of the converter was derived, and algorithms for testing this condition as well as constructing the *maximally permissive* converter were obtained. Also, heuristic techniques with polynomial complexity for converter synthesis were presented. Rudie-Wonham [26] have also applied the supervisory control techniques for deriving a missing portion (either sender or receiver portion) of the protocol assuming some knowledge of it. This problem is quite similar to that of protocol conversion except it does not consider the progress specification.

In this paper we show that the protocol conversion problem can be reduced to the problem of designing a nonblocking supervisor by suitably defining a *family* of marked languages. A supervisor is said to be nonblocking with respect to a marking if every generated trace of the controlled system can also be extended to a marked trace of the controlled system. We show that the protocol conversion system meets the progress constraint if and only if it is nonblocking with respect to a family of markings—one for each external event. In case the given service specification cannot be implemented, we provide techniques to *minimally* alter it so that it becomes implementable. Alterations by way of both minimal relaxation and minimal constraining the given specification are considered. This complements the recent work of Takai-Takae-Kodama [28] on the problem of minimally altering the service specifications.

Finally we obtain more efficient converters that are *minimally permissive*, which avoid execution of spurious internal events. The algorithm presented here computes such a supervisor *on-line*. The complexity of the algorithm is *linear* in the number of states in the product of the states in the system and the automaton representation of the specification. This complements the earlier work [17, 18] where an off-line synthesis procedure for a maximally permissive supervisor was presented.

2 Notation and Preliminaries

We use Σ to denote the universe of events. Σ^* denotes the set of all finite length event sequences, called *traces*, including the zero length trace, denoted ϵ. A subset of Σ^*, i.e., a collection of traces, is called a *language*. For a language H, the notation $pr(H)$, called the *prefix closure* of H, is the set of all prefixes of traces in H. H is said to be *prefix closed* if $H = pr(H)$. The notation $pr(s)$ denotes the set of all prefixes of a trace $s \in \Sigma^*$. Given $s \in \Sigma^*$ and a subset of events $\hat{\Sigma} \subseteq \Sigma$, the *projection* of s on $\hat{\Sigma}$, denoted $s\uparrow\hat{\Sigma}$, is the trace obtained by erasing the events not belonging to $\hat{\Sigma}$ from s. Given $H_1, H_2 \subseteq \Sigma^*$, H_1 *quotient* H_2, defined as

$$H_1/H_2 := \{s \in \Sigma^* \mid \exists t \in H_2 : st \in H_1\},$$

denotes the set of traces for which there exists an extension in H_2 so that the extended trace is in H_1, i.e., these are the traces obtained by the removal of an H_2-suffix from an H_1-trace.

State machines [9] are used for representing untimed behavior of discrete event systems (such as protocol conversion systems) as well as for representing qualitative or logical specifications (such as user service specifications). Formally, a state machine P is a quadruple $P := (X, \Sigma, \alpha, x_0)$, where X denotes the set of states, Σ denotes the finite set of events, $\alpha : X \times (\Sigma \cup \{\epsilon\}) \to 2^X$ is the partial transition function, and $x_0 \in X$ is the initial state. We view a discrete event system to be only a generator of event sequences, and hence do not define the notion of marked set of states. In contrast, the specification imposes a marking on the generated traces and marks those traces that result in the accomplishment of the specification.

A triple $(x, \sigma, x') \in X \times (\Sigma \cup \{\epsilon\}) \times X$ is called a *transition* of P if $x' \in \alpha(x, \sigma)$. The transition function is extended from events to traces $\alpha : X \times \Sigma^* \to X$ in a natural way. The *generated language* of P is the set of all traces that it can execute starting from its initial state, i.e.,

$$L(P) := \{s \in \Sigma^* \mid \alpha(x_0, s) \neq \emptyset\}.$$

P is called a *deterministic* state machine if the transition function is a partial function of the form: $\alpha : X \times \Sigma \to X$. We only consider systems with deterministic state machine representation in this paper.

Synchronous composition [8] of deterministic state machines is used to represent concurrent operation of component systems, such as the system with mismatched set of protocols and the protocol converter. Synchronous composition results in a restricted behavior and thus can be also used for supervisory control of a discrete event plant. Given two deterministic state machines $P := (X_P, \Sigma_P, \alpha_P, x_{0,P})$ and $Q := (X_Q, \Sigma_Q, \alpha_Q, x_{0,Q})$, composition of P and Q, denoted $P\|Q := (X, \Sigma, \alpha, x_0)$, is defined as: $X := X_P \times X_Q$, $\Sigma := \Sigma_P \cup \Sigma_Q$, $x_0 := (x_{0,P}, x_{0,Q})$, and for each $x = (x_P, x_Q) \in X$ and $\sigma \in \Sigma$:

$$\alpha(x, \sigma) := \begin{cases} (\alpha_P(x_P, \sigma), \alpha_Q(x_Q, \sigma)) & \text{if } \alpha_P(x_P, \sigma), \alpha_Q(x_Q, \sigma) \text{ defined}, \sigma \in \Sigma_P \cap \Sigma_Q \\ (\alpha_P(x_P, \sigma), x_Q) & \text{if } \alpha_P(x_P, \sigma) \text{ defined}, \sigma \in \Sigma_P - \Sigma_Q \\ (x_P, \alpha_Q(x_Q, \sigma)) & \text{if } \alpha_Q(x_Q, \sigma) \text{ defined}, \sigma \in \Sigma_Q - \Sigma_P \\ \text{undefined} & \text{otherwise} \end{cases}$$

Thus when P and Q are composed, their common events occur synchronously, while the other events occur asynchronously. Note that the common events that are not feasible in either of the systems are prevented from occurring or get *disabled* in the composition of the two systems, which is how synchronization achieves control. i.e., enablement/disablement of events. The generated language of the composition is given by:

$$L(P\|Q) = \{s \in \Sigma^* \mid s\uparrow\Sigma_P \in L(P), s\uparrow\Sigma_Q \in L(Q)\}.$$

Note that when $\Sigma_P = \Sigma_Q = \Sigma$, then $L(P\|Q) = L(P) \cap L(Q)$ since all events must occur synchronously. Also note that although the state set for $P\|Q$ is $X_P \times X_Q$, many of the states remain unreachable in $P\|Q$. We adopt the convention that by writing $P\|Q$ we mean its *reachable* or *trim* component [9].

In supervisory control, synchronous composition of the supervisor and the uncontrolled plant is used as one possible control mechanism. Both the supervisor S, and the uncontrolled plant G are modeled as state machines with the same event sets, say Σ. The generated language of the controlled system equals $L(G\|S) = L(G) \cap L(S)$; and a suitably defined sublanguage of it, denoted $L_m(G\|S)$, contains the *marked* traces of the controlled behavior representing those traces execution of which accomplishes certain tasks of interest. $L_m(G\|S)$ is called the *marked language* of the controlled system. and the supervisor S is said to be *nonblocking* with respect to the marking of traces if $L(G\|S) = pr(L_m(G\|S))$, i.e., if each generated trace is a prefix of some marked trace so that it can be extended to a trace accomplishing a task of interest.

A certain sublanguage $H \subseteq L(G)$ represents the desired behavior of the plant, and the control objective is to design a nonblocking supervisor S such that the marked language of the controlled plant equals H. The supervisor to be designed has limited capabilities in the sense that (i) it cannot prevent the occurrence of certain *uncontrollable* events, and (ii) it can only observe the occurrence of certain ***observable*** events. Let $\Sigma_u \subseteq \Sigma$ denote the set of uncontrollable events and $\Sigma_o \subseteq \Sigma$ denote the set of observable events. Then due to limited control and observation capability of a supervisor, only those languages can be achieved as controlled behavior that are [25, 20]:

Controllable: $pr(H)\Sigma_u \cap L(G) \subseteq pr(H)$.

Observable: $\forall s, t \in pr(H), \sigma \in \Sigma : s\uparrow\Sigma_o = t\uparrow\Sigma_o, s\sigma \in pr(H), t\sigma \in L(G) \Rightarrow t\sigma \in pr(H)$.

The controllability condition requires that extension of any prefix of H by an uncontrollable event that is feasible in the plant should also be a prefix of H. This is because the occurrence of uncontrollable events cannot be prevented. A pair of traces is called *indistinguishable* if each trace in the pair has identical projection on the set of observable events. The observability condition requires that extensions of a pair of indistinguishable prefixes of H by a common feasible event should either both or neither be prefixes of H. This is because the same set of events must be enabled/disabled following an indistinguishable pair of traces.

Note that H is controllable (respectively, observable) if and only if $pr(H)$ is controllable (respectively, observable). It is known that controllability is preserved under union so that a unique maximal controllable sublanguage, called the *supremal controllable sublanguage* of a given language exists [24]; however, observability is not preserved under union, so no unique maximal observable sublanguage exists [20]. Hence sometimes *normal* sublanguages instead of observable sublanguages are considered [20]:

Normal: $\forall s, t \in \Sigma^* : s\uparrow\Sigma_o = t\uparrow\Sigma_o, s \in pr(H), t \in L(G) \Rightarrow t \in pr(H)$.

Normality requires that traces of G that are indistinguishable from a prefix of H must themselves be prefixes of H, and it is a stronger condition than observability. Furthermore it is preserved under union so that the *supremal normal sublanguage* of a given language exists [20, 2, 16].

It is also known that when the set of controllable events is a subset of that of the observable events ($\Sigma - \Sigma_u \subseteq \Sigma_o$), then controllability together with observability is equivalent to controllability together with normality [15]. This, for instance, is the case in the setting of protocol converters—a converter can observe its input and output events and can control only its output events. Consequently, a unique maximally permissive supervisor that achieves the supremal prefix closed, controllable, and normal sublanguage of the specification language can be achieved.

3 Progress Constraint as Nonblocking

We are interested in designing a supervisor (a converter in case of protocol conversion) such that the controlled system "implements" a partial service specification specified on the external set of events through which the system interacts with the user at the next higher level. We use $\Sigma_e \subseteq \Sigma$ to denote the set of external events, and $\Sigma_i := \Sigma - \Sigma_e$ to denote the set of internal events.

Definition 1 Given a system P and a prefix closed partial specification $K \subseteq \Sigma_e^*$, P *implements* K if the following hold:

Safety: $L(P)\uparrow\Sigma_e \subseteq K$.

Progress: $\forall s \in L(P), \sigma \in \Sigma_e : (s\uparrow\Sigma_e)\sigma \in K \Rightarrow s \in L(P)/\Sigma_i^*\sigma$.

Note that the containment in the safety property can be replaced by equality since the reverse containment follows from progress. Safety requires that each generated trace of P should correspond to a trace of the specification, i.e., no "illegal" traces should occur in P. Since K is a partial specification (defined only on the external event set), there may exist more than one trace of P that correspond to the same trace of K. Progress requires that if an external event is possible after a trace of K, then this external event should also be possible "eventually", i.e., after occurrence of zero or more internal events following each corresponding trace of P. The definition of "implements" given above was introduced in [17, 18] and is equivalent to that first given in [3] but is stated differently for simplicity.

The progress constraint requires that the controlled system should never block an external event whenever the service specification does not block it. This suggests that the progress constraint could be characterized as a nonblocking property of the controlled system by suitably defining the marked traces of the system. We show in this section that this is indeed true. However, this requires that a family of marked languages be defined—one for each external event—and that the controlled system be nonblocking with respect to each of the markings.

Definition 2 Consider a system P, a prefix closed partial specification $K \subseteq \Sigma_e^*$, and an external event $\sigma \in \Sigma_e$. $s \in L(P)$ is said to be *σ-terminal* if the last event of s is σ; it is said to be *σ-blocking in K* if $(s\uparrow\Sigma_e) \in K$ and $(s\uparrow\Sigma_e)\sigma \notin K$. Define

$$L_T^\sigma(P) := \{s \in L(P) \mid s \text{ is } \sigma\text{-terminal}\}; \quad L_B^\sigma(P) := \{s \in L(P) \mid s \text{ is } \sigma\text{-blocking in } K\}.$$

The *σ-marked language* of P is defined as:

$$L_m^\sigma(P) := \{s \in L_T^\sigma(P) \cup L_B^\sigma(P) \mid \forall t \in pr(s), \hat{\sigma} \in \Sigma_e : (t\uparrow\Sigma_e)\hat{\sigma} \in K \Rightarrow t \in L(P)/\Sigma_i^*\hat{\sigma}\}.$$

$s \in L(P)$ is said to be *σ-marked* if $s \in L_m^\sigma(P)$.

In other words, a trace in P is said to be σ-marked if either it is σ-terminal (so that it does not block σ) or it is σ-blocking in K (so that the specification itself blocks σ), *and* whenever the specification does not block an external event $\hat{\sigma}$ at a certain prefix t of s, then this prefix can be extended by internal events followed by $\hat{\sigma}$. In the following theorem we show that the above definition of marking is adequate, as it characterizes progress in terms of nonblockingness. From now on, we will use state machine G to denote an uncontrolled system (representing for example the mismatch set of protocols) with event set Σ, $K \subseteq \Sigma_e^*$ to denote a prefix closed partial service specification, and S to denote a supervisor that controls G by operating in complete synchrony with it and restricts its behavior to that of $G\|S$. The objective of supervision is that $G\|S$ should implement K. We assume without loss of generality that $K \subseteq L(G)\uparrow\Sigma_e$ (otherwise replace K by $K \cap L(G)\uparrow\Sigma_e$).

Definition 3 Given G and $K \subseteq \Sigma_e^*$, a supervisor S is said to be *Σ_e-nonblocking* if for each $\sigma \in \Sigma_e$ it is nonblocking with respect to each σ-marking, i.e., $pr(L_m^\sigma(G\|S)) = L(G\|S)$.

Theorem 1 Given G and $K \subseteq \Sigma_e^*$, a supervisor S achieves progress, i.e., $L(G\|S)$ satisfies progress if and only if S is Σ_e-nonblocking.

In [17, Theorem 2] it was shown that the existence of a supervisor implementing a given service specification is equivalent to the existence of a nonempty prefix closed sublanguage of the uncontrolled plant language that is controllable, normal, and satisfies safety and progress. Such a language was defined to be a *converter* language [17, Definition 2]. In the following corollary we provide an alternative characterization of a converter language.

Definition 4 Given G and $K \subseteq \Sigma_e^*$, a prefix closed sublanguage $H \subseteq L(G)$ is said to be *Σ_e-nonblocking* if for each $\sigma \in \Sigma_e$, $H = pr(H_m^\sigma)$, where

$$H_m^\sigma := H \cap \{s \in L_T^\sigma(G) \cup L_B^\sigma(G) \mid \forall t \in pr(s), \hat{\sigma} \in \Sigma_e : (t\uparrow\Sigma_e)\hat{\sigma} \in K \Rightarrow t \in H/\Sigma_i^*\hat{\sigma}\}.$$

Then the following corollary follows from the previous theorem:

Corollary 1 Given G and $K \subseteq \Sigma_e^*$, a prefix closed sublanguage $H \subseteq L(G)$ is a converter language if and only if it is controllable, normal, satisfies safety, and is Σ_e-nonblocking.

Let $Conv(G, K)$ denote the set of all converter sublanguages of G with respect to the specification K. It was shown ([17, Corollary 1]) that a unique maximal converter sublanguage, denoted $supConv(G, K)$, exists. This implies that the Σ_e-nonblockingness property of languages is preserved under union. Given any partial specification $R \subseteq \Sigma_e^*$, we use

$$R_G^{-1} := \{s \in L(G) \mid s\uparrow\Sigma_e \in R\}$$

to denote the *inverse projection of R on $L(G)$*, i.e., it is the set of traces of G whose projection on external events is in R. We show in the next theorem that $supConv(G, K)$ equals the supremal prefix closed controllable, normal, and Σ_e-nonblocking sublanguage of K_G^{-1}.

Theorem 2 Given G and $K \subseteq \Sigma_e^*$, $supConv(G, K)$ equals the supremal prefix closed, controllable, normal, and Σ_e-nonblocking sublanguage of K_G^{-1}.

It follows from the above theorem that the supremal converter language can be computed by computing the supremal controllable, normal, and Σ_e-nonblocking sublanguage of K_G^{-1}. This fact can be used to simplify the algorithm for the computation of $supConv(G, K)$ given in [17, Algorithm 3]. Note that an automaton that generates K_G^{-1} can be easily obtained by taking an automaton generator for K, adding self-loops on each internal event at its each state, and synchronously composing the resulting automaton with G.

4 Minimally Altered Service Specifications

The existence of a supervisor implementing a given partial service specification is equivalent to the supremal converter language being nonempty. However, a situation may arise in which the supremal converter language is empty and thus no supervisor exists. Then we could consider the possibility of *minimally* altering the service specification so that a supervisor could be designed for this altered service specification. There are two ways of altering the service specification—by relaxing or weakening it by enlarging it, or by constraining or strengthening it by restricting it.

We consider the design based on relaxing the service specification first. If the supremal prefix closed, controllable, normal, and Σ_e-nonblocking sublanguage of K_G^{-1} is empty, a natural choice for the supervisor is one which achieves a minimal prefix closed, controllable, normal, and Σ_e-nonblocking superlanguage of K_G^{-1}. That such a minimal exists follows from the fact that given any *chain* of superlanguages of K_G^{-1} such that each language in the chain is prefix closed, controllable, normal, and Σ_e-nonblocking , their intersection is also prefix closed, controllable, normal, and Σ_e-nonblocking superlanguage of K_G^{-1}. Let $(K_G^{-1})^{\downarrow}$ be one such minimal superlanguage. In the next theorem we show that the projection of $(K_G^{-1})^{\downarrow}$ on the set of external events is a minimally relaxed service specification which can be implemented via supervision.

Theorem 3 Suppose G and $K \subseteq \Sigma_e^*$ are such that $supConv(G, K) = \emptyset$. Let $(K_G^{-1})^{\downarrow}$ be a minimal prefix closed, controllable, normal, and Σ_e-nonblocking superlanguage of K_G^{-1}. Then for any $\hat{K} \subset [(K_G^{-1})^{\downarrow}]\uparrow\Sigma_e$ such that $\hat{K} \supseteq K$, $supConv(G, \hat{K}) = \emptyset$.

Remark 1 Theorem 3 states that if K cannot be implemented by supervision, but a minimal prefix closed controllable, normal, and Σ_e-nonblocking superlanguage of K_G^{-1}, denoted $(K_G^{-1})^{\downarrow}$ exists, then its projection on the external event set is a minimally relaxed service specification which can be implemented by supervision (any other relaxed service specification $\hat{K} \supseteq K$ that is contained in $[(K_G^{-1})^{\downarrow}]\uparrow\Sigma_e$ can also not be implemented).

It is easy to see that the Σ_e-nonblocking property is not preserved under intersection; consequently, a unique minimal prefix closed, controllable, normal, and Σ_e-nonblocking superlanguage of K_G^{-1} does not exist [28, Example 4]. However, the Σ_e-nonblocking property could be replaced by the stronger property of Σ_e-normality, which is preserved under intersection.

Definition 5 Given G, and the set of external events $\Sigma_e \subseteq \Sigma$, a language $H \subseteq L(G)$ is said to be *Σ_e-normal* if

$$\forall s, t \in \Sigma^* : s\uparrow\Sigma_e = t\uparrow\Sigma_e, s \in pr(H), t \in L(G) \Rightarrow t \in pr(H).$$

In other words, Σ_e-normality is same as normality with the observable event set being the set Σ_e. Clearly, Σ_e-normality is preserved under intersection and union. In the next theorem we show that Σ_e-normality together with the fact that $\Sigma_e \subseteq \Sigma_u$ implies Σ_e-nonblockingness.

Theorem 4 Given G and $K \subseteq \Sigma_e^*$, suppose a prefix closed sublanguage $H \subseteq L(G)$ is controllable, and Σ_e-normal. Then H is Σ_e-nonblocking.

Remark 2 It follows from the previous theorem that one way to relax or enlarge a given service specification K (which itself cannot be implemented via supervision) is to consider the infimal prefix closed, controllable, normal, and Σ_e-normal superlanguage of K_G^{-1} and then take its projection on the external events. This can be algorithmically computed using the existing computational techniques from supervisory control theory (refer for example [15]). Such a service specification need not be a minimally relaxed one, as we have replaced the Σ_e-nonblockingness by the stronger property of Σ_e-normality, but it is simpler to compute. The computation of a minimally relaxed specification is explained in Remark 1.

In the remainder of this section we consider design based on minimally constraining the service specification which itself cannot be implemented. If $supConv(G, K) = \emptyset$, then one could consider implementing a maximal sublanguage $\hat{K} \subseteq K$ for which $supConv(G, \hat{K})$ is nonempty. It is easy to show that such maximals are well defined since any given *chain* of increasing sub-specifications:

$$\{K_i \subseteq K \mid supConv(G, K_i) \neq \emptyset\}$$

has the property that $supConv(G, \cup_i K_i) \neq \emptyset$. However, a unique maximal sub-specification for which supremal converter sublanguage is nonempty need not exist. This is because the progress property is not preserved—given $K_i \subseteq K$ such that $supConv(G, K_i) \neq \emptyset$, there need not exist a sublanguage which satisfies progress with respect to $\cup_i K_i$ [28, Example 3].

In order to be able to *compute* a service sub-specification which can be implemented under supervision we replace the property of Σ_e-nonblockingness with the stronger property of Σ_e-normality. Given any sub-specification $\hat{K} \subseteq K$, we use $supPCNSN_e(G, \hat{K})$ to denote the supremal prefix closed, controllable, normal, safe, and Σ_e-normal sublanguage of $\hat{K}_G^{-1}$. Then since Σ_e-normality is stronger than Σ_e-nonblockingness, $sup\,Conv(G, \hat{K}) \supseteq supPCNSN_e(G, \hat{K})$, and nonemptiness of $supPCNSN_e(G, \hat{K})$ is only a sufficient condition for the existence of a supervisor that implements $\hat{K}$.

Theorem 5 Given G and $K \subseteq \Sigma_e^*$, suppose the set of sub-specifications:

$$\{K_i \subseteq K \mid supPCNSN_e(G, K_i) \neq \emptyset\} \neq \emptyset.$$

Then $supPCNSN_e(G, \cup_i K_i) \neq \emptyset$.

Remark 3 It follows from the last theorem that whenever there exists a sub-specification for which the corresponding supremal prefix closed, controllable, normal, safe, and Σ_e-normal sublanguage is nonempty, then there exists a unique maximal such sub-specification. However, this sub-specification is not necessarily a minimally constrained sub-specification that is implementable under supervision, since we have replaced the Σ_e-nonblockingness by the stronger property of Σ_e-normality.

In the next theorem we provide a way to compute the unique supremal sub-specification $\hat{K} \subseteq K$, such that $supPCNSN_e(G, \hat{K}) \neq \emptyset$ by showing that it can be computed by computing the supremal prefix closed, controllable, normal, and Σ_e-normal sublanguage of K_G^{-1}, denoted $supPCNN_e(K_G^{-1})$, and next projecting it on to the set of external events.

Theorem 6 Given G and $K \subseteq \Sigma_e^*$, let $K^\dagger \subseteq K$ be the supremal sub-specification for which $supPCNSN_e(G, K^\dagger) \neq \emptyset$. Then $K^\dagger = [supPCNN_e(K_G^{-1})]\uparrow\Sigma_e$.

Remark 4 Theorem 6 provides a formula for $K^\dagger \subseteq K$, the supremal sub-specification for which the supremal prefix closed, controllable, normal, safe, and Σ_e-normal sublanguage is nonempty, and hence $K^\dagger$ is implementable under supervision. However, since we have replaced the property of Σ_e-nonblockingness by Σ_e-normality, $K^\dagger$ need not be a minimally constrained sub-specification of K that is implementable under supervision. But the result of Theorem 6 suggests a computation procedure for such a minimally constrained sub-specification: First compute a minimal prefix closed, controllable, normal, and Σ_e-nonblocking sublanguage of K_G^{-1}, and next consider its projection on the set of external events.

5 Minimally Permissive Converters

The supervisor which achieves $sup\,Conv(G, K)$ as the controlled behavior is maximally permissive. This converter is unique and is easily computed off-line. However, by virtue of being maximally permissive it allows execution of spurious traces of internal events that may be avoided. In order to prevent execution of such spurious traces, it is desirable to have a minimally permissive supervisor that achieves a minimal converter language as the controlled system behavior. But the set of converter languages $Conv(G, K)$ is not closed under intersection (both safety and progress are not preserved under intersection), consequently a unique minimal converter language does not always exist, and furthermore it may not be possible to compute one *off-line* (a minimal converter language need not be regular even though the system behavior $L(G)$ and the service specification K are given to be regular languages). In this section we present an *on-line* technique to compute a minimal converter language.

On-line control computation algorithm for control under partial observation was first presented in [7], and later generalized in [6] for the computation of maximally permissive supervisors. The computational complexity of these algorithms is linear in the number of states in the product of plant and specification automata. Our algorithm has the identical complexity but in contrast it computes a *minimally permissive* supervisor that not only obtains a controllable and observable behavior but also that satisfies safety and progress.

The on-line control computation algorithm computes a minimal set of controllable events to be enabled after the occurrence of each observable event. For this we first refine the system with respect to the specification. Let $S := (Y, \Sigma_e, \beta, y_0)$ be a trim automaton representing the specification, i.e., $L(S) = K$. Let $\overline{S} := (Y \cup \{d\}, \Sigma_e, \overline{\beta}, y_0)$ be the automaton obtained by "completing" the transition function of S by introducing a dump state d in S and adding transitions from each state to the dump state on those events that are not defined at that state, i.e.,

$$\forall y \in Y \cup \{d\}, \sigma \in \Sigma_e : \overline{\beta}(y, \sigma) := \begin{cases} \beta(y, \sigma) & \text{if } \beta(y, \sigma) \text{ is defined} \\ d & \text{otherwise} \end{cases}$$

The system is refined with respect to the specification by composing G and $\overline{S}$, i.e., by computing $G\|\overline{S} := (Z, \Sigma, \gamma, z_o)$. Then it is easy to see that $L(G\|\overline{S}) = L(G)$ and states of $G\|\overline{S}$ with second coordinate the dump state are reached by those traces in $L(G) = L(G\|\overline{S})$ that violate the specification. Such states of $G\|\overline{S}$ are called *illegal*, denoted $Z_i \subseteq Z$, and the remaining states are called *legal*, denoted $Z_l := Z - Z_i$. The set of states from where it is possible to reach the illegal states in zero or more uncontrollable transitions is called the *bad* states, denoted $Z_b \supseteq Z_i$, and the remaining states are called the good states, denoted $Z_g := Z - Z_b \subseteq Z_l$.

Given $\hat{\Sigma} \subseteq \Sigma$ and $\hat{Z} \subseteq Z$, four reachability operations are defined over the refined plant $G\|\overline{S}$:

Reach: $R_{\hat{\Sigma}}(\hat{Z})$ denotes the set of states reachable from $\hat{Z}$ on traces in $\hat{\Sigma}^*$

Restricted reach: $R_{-\hat{\Sigma}}(\hat{Z})$ denotes the set of states reachable from $\hat{Z}$ on traces in $\hat{\Sigma}^*$ without visiting states in Z_b

Unobservable reach [6]: $UR_{\hat{\Sigma}}(\hat{Z})$ denotes the set of states reachable from $\hat{Z}$ on traces in $(\hat{\Sigma} - \Sigma_o)^*$

Extended unobservable reach [6]: $UR_{+\hat{\Sigma}}(\hat{Z})$ denotes the set of states reachable from $UR_{\hat{\Sigma}}(\hat{Z})$ on zero or single event in $\hat{\Sigma} \cap \Sigma_o$

The algorithm maintains in variable $\hat{Z} \subseteq Z$ the set of possible system states immediately after the occurrence on an observable event. $\hat{Z}$ is initialized to $\{z_0\}$ when no observation has taken place. In the variable $ACT \subseteq \Sigma$, where $ACT \supseteq \Sigma_u$, the algorithm computes a minimal set of events to be enabled immediately after the occurrence of an observable event. The set $Elist = \Sigma - \Sigma_u$ contains an ordered list of controllable events. In order to arrive at a final value for the variable ACT, a temporary variable $\hat{\Sigma} \subseteq \Sigma$, which is initialized to Σ_u, is incrementally enlarged by adding appropriate controllable events one at a time till the set of enabled events in $\hat{\Sigma}$ is adequate enough to not block any feasible external event. To do this first the set of feasible external events of the specification when the corresponding system state is $z \in UR_{\hat{\Sigma}}(\hat{Z})$—a state system could reach unobservably if the starting states are $\hat{Z}$ and the enabled events are $\hat{\Sigma}$, is computed in the variable $\hat{\Sigma}_e(z) \subseteq \Sigma_e$. Next the extended unobservable reach $UR_{+(\hat{\Sigma}\cap\Sigma_i)}(z)$ is computed; this the set of states the system could reach immediately before or after an observation by executing only internal events when the set of starting states are $\hat{Z}$ and the set of enabled events are $\hat{\Sigma}$. If the restricted reach on internal events of these states, $R_{-\Sigma_i}[UR_{+(\hat{\Sigma}\cap\Sigma_i)}(z)]$, allows $\hat{\Sigma}_e(z)$, i.e., if $\hat{\Sigma}_e(z)$ is defined at those states that are reached from executions of internal events starting from states in $UR_{+(\hat{\Sigma}\cap\Sigma_i)}(z)$ and without ever visiting a bad state, then the variable ACT is set to the value that temporary variable $\hat{\Sigma}$ holds; otherwise, $\hat{\Sigma}$ is incremented by another controllable event and the steps are repeated till the list of controllable events is exhausted. If at any time during these steps the set $UR_{+\hat{\Sigma}}(\hat{Z})$ is not contained in the set of good states, then the last controllable event added to $\hat{\Sigma}$ is removed from it.

Algorithm 1 Given $G := (X, \Sigma, \alpha, x_0)$ and a trim generator $S := (Y, \Sigma, \beta, y_0)$ for $K \subseteq \Sigma_e^*$, compute the refined plant $G\|\overline{S} := (Z, \Sigma, \gamma, z_0)$.

```
1. Initialization:
   $\hat{Z} := \{z_0\}$; if $\hat{Z} \not\subseteq Z_g$, then ERROR and STOP
2. Control Computation:
   (a) $\hat{\Sigma} := \Sigma_u, pt := 0$
   (b) For each $z = (x, y) \in UR_{\hat{\Sigma}}(\hat{Z})$, compute
           $\hat{\Sigma}_e(z) := \{\sigma \in \Sigma_e \mid \exists(y, \sigma) \text{ defined}\}; \quad Z^*(z) := R_{-\Sigma_i}[UR_{+(\hat{\Sigma}\cap\Sigma_i)}(z)].$
   (c) If $\hat{\Sigma}_e(z)$ defined at $Z^*(z)$ for each $z \in UR_{\hat{\Sigma}}(\hat{Z})$, then $ACT := \hat{\Sigma}$ and Goto 3;
       Else if $pt = |Elist|$, then ERROR and STOP; else $pt := pt + 1, \hat{\Sigma} := \hat{\Sigma} \cup \{Elist.pt\}$
   (d) If $UR_{+\hat{\Sigma}}(\hat{Z}) \subseteq Z_g$, then Goto 2(b);
       Else if $pt \neq |Elist|$, then $\hat{\Sigma} := \hat{\Sigma} - \{Elist.pt\}, pt := pt + 1, \hat{\Sigma} := \hat{\Sigma} \cup \{Elist.pt\}$, Goto 2(d); else
       ERROR and STOP
3. Update state-estimate:
   Wait for $\sigma \in \Sigma_o$ to occur, update $\hat{Z} := \gamma(UR_{ACT}(\hat{Z}), \sigma)$, and Goto 2
```

In the following theorem we prove the correctness of Algorithm 1.

Theorem 7 Suppose $supConv(G,K) \neq \emptyset$, and let $H \subseteq L(G)$ be the behavior of the controlled system under the control of supervisor computed using Algorithm 1. Then H is a minimal converter language.

Remark 5 In Algorithm 1, the computation of a control action after any observation requires only the evaluation of various reachability operations. Hence the computational complexity at each step is linear in the number of states in the refined plant. The on-line technique is not only computationally efficient (the off-line computation of supervisor in [17, 18] has an exponential complexity), it is also more effective since it is minimally permissive and thereby avoids execution of spurious internal traces. By changing the order of controllable events in the list *Elist*, it is possible to obtain different minimally permissive supervisors.

6 Conclusion

We have studied the problem of designing a supervisor that *implements* a partial specification. Such problems arise in systems that possess a hierarchically layered architecture such as communication protocol. We have transformed the problem of "implementability" of a *partial* specification to that of achieving an appropriate *complete* specification which is required to be nonblocking with respect to an appropriately defined family of markings. Nonblockingness with respect to a *family* of markings is itself an interesting supervisory control problem, and is investigated further in [4]. The problem of minimally altering a given specification to make it implementable under supervision has also been addressed here. We have given constructive techniques for minimally relaxing and also for minimally constraining a given specification so that it becomes implementable. The notion of Σ_e-normality plays an important role in this regard. Finally we presented an on-line control computation algorithm of linear complexity that implements a given service specification effectively by only executing a minimal converter language as the controlled system behavior.

Acknowledgment

We thank Prof. S. Lafortune for suggesting representing the progress constraint as a nonblocking constraint.

References

[1] G. V. Bochmann. Deriving protocol converters for communication gateways. *IEEE Transactions on Communications*, 38(9):1298–1300, 1990.

[2] R. D. Brandt, V. K. Garg, R. Kumar, F. Lin, S. I. Marcus, and W. M. Wonham. Formulas for calculating supremal controllable and normal sublanguages. *Systems and Control Letters*, 15(8):111–117, 1990.

[3] K. Calvert and S. S. Lam. Formal methods of protocol conversion. *IEEE Journal on Selected Areas in Communication*, 8(1):127–142, January 1990.

[4] M. Fabian and R. Kumar. Mutually nonblocking supervisory control of discrete event systems. In *1997 IEEE Conference on Decision and Control*, San Diego, CA, 1997. Submitted.

[5] P. E. Green. Protocol conversion. *IEEE Transactions on Communications*, COM-34(3):257–268, March 1986.

[6] N. B. Hadj-Alouane, S. Lafortune, and F. Lin. Centralized and distributed algorithm for on-line synthesis of maximal control policies under partial observation. *Discrete Event Dynamical Systems: Theory and Applications*, 6(41):379–427, 1996.

[7] M. Heymann and F. Lin. On-line control of partially observed discrete event systems. *Discrete Event Dynamical Systems: Theory and Applications*, 4(3):221–236, July 1994.

[8] C. A. R. Hoare. *Communicating Sequential Processes.* Prentice Hall, Inc., Englewood Cliffs, NJ, 1985.

[9] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley, Reading, MA, 1979.

[10] M.-C. Hsu and S.-C. Chang. A supervisory control theoretic approach for missing module construction. In *Proccedinds of 1996 International Workshop on Discrete Event Systems*, pages 108–113, Edinburgh, UK, August 1996.

[11] K. Inan. Supervisory control and formal methods for distributed systems. In *Discrete Event Systems: Modeling and Control (Proceedings of WODES 1992)*, pages 29–41. Birkhauser-Verlag-Basel, 1993.

[12] K. Inan. Supervisory control: Theory and application to the gateway synthesis problem. In *Belgian-French-Netherlands Summer School on Discrete Event Systems*, page 25 pages. Spa, Belgium, 1993.

[13] K. Inan. Nondeterministic supervision under partial observations. In Guy Cohen and Jean-Pierre Quadrat, editors, *Lecture Notes in Control and Information Sciences 199*, pages 39–48. Springer-Verlag, New York, 1994.

[14] D. M. Kristol, D. Lee, A. N. Netravali, and K. Sabnani. A polynomial algorithm for gateway generation from formal specifications. *IEEE Transactions on Networking*, 1(2):217–229, 1993.

[15] R. Kumar and V. K. Garg. *Modeling and Control of Logical Discrete Event Systems.* Kluwer Academic Publishers, Boston, MA, 1995.

[16] R. Kumar, V. K. Garg, and S. I. Marcus. On controllability and normality of discrete event dynamical systems. *Systems and Control Letters*, 17(3):157–168, 1991.

[17] R. Kumar, S. Nelvagal, and S. I. Marcus. Design of protocol converters: A discrete event systems approach. In *Proceedings of 1996 International Workshop on Discrete Event Systems*, pages 7–12, Edinburgh, UK, August 1996.

[18] R. Kumar, S. Nelvagal, and S. I. Marcus. A discrete event systems approach for protocol conversion. *Discrete Event Dynamical Systems: Theory and Applications*, 1996. Accepted.

[19] Simon S. Lam. Protocol conversion. *IEEE Transactions on Software Engineering*, 14(3):353–362, March 1988.

[20] F. Lin and W. M. Wonham. On observability of discrete-event systems. *Information Sciences*, 44(3):173–198, 1988.

[21] K. Okumura. A formal protocol conversion method. *Proceedings ACM SIGCOMM*, pages 30–37, 1986.

[22] A. Overkamp. Supervisory control using failure semantics and partial specifications. *IEEE Transactions on Automatic Control*, 41, 1996. To Appear.

[23] M. Rajagopal and R. E. Miller. Synthesizing a protocol converter from executable protocol traces. *IEEE Transactions on Computers*, 40(4):487–499, 1991.

[24] P. J. Ramadge and W. M. Wonham. On the supremal controllable sublanguage of a given language. *SIAM Journal of Control and Optimization*, 25(3):637–659, 1987.

[25] P. J. Ramadge and W. M. Wonham. Supervisory control of a class of discrete event processes. *SIAM Journal of Control and Optimization*, 25(1):206–230, 1987.

[26] K. Rudie and W. M. Wonham. Supervisory control of communicating processes. In L. Logrippo, R. L. Robert, and H. Ural, editors, *Protocol Specification, Testing and Verification*, pages 243–257. Elsevier Science Publishers, North-Holland, 1990.

[27] A. Uday Shankar. Modular design principles for protocols with an application to the transport layer. *Proceedings of The IEEE*, 79(12):1687–1706, 1991.

[28] S. Takai, A. Takae, and S. Kodama. The extremal languages arising in supervisory control for service specifications. In *Proceedings of 1996 IEEE Conference on Decision and Control*, Kobe, Japan, December 1996.

[29] A. S. Tanenbaum. *Computer Networks.* Prentice Hall, Inc., Englewood Cliffs, NJ, 1990.

Performance Opportunity for Flexible Workers [1]

Mark P. Van Oyen, Esma Senturk-Gel, and Wallace J. Hopp

Department of Industrial Engineering and Management Sciences
Northwestern University, Evanston, IL 60208-3119

Abstract

We formulate an idealistic model for flexible workers in a series queueing network. The control issue is how to assign each worker to jobs/stations over time. Under general assumptions and complete worker flexibility, we prove that a simple policy, which we call the expedite policy, minimizes along each sample path the cycle time (delay) for each job. Therefore, the expedite policy also minimizes work in process and maximizes throughput along every sample path for any time t. We also consider the optimal static allocation of effort (workers) and use this to compute the performance improvement opportunity achievable using flexible workers. We identify conditions under which the opportunity of flexible workers is significant. Finally, we briefly examine closed queueing network models, which provide excellent intuition regarding the benefits of flexible workers.

1. Introduction

Much research found in the literature has focused on highly rigid models of production based on the concept of division of labor and the production line. This is not surprising for two reasons. First, the industrial revolution replaced the "craftsman" approach to production with a paradigm based on standardization and division of labor. Secondly, both mathematical and conceptual models of production are relatively complex, even for the case of rigid production systems. For systems with flexible workers that can be allocated anywhere in the system, the models quickly become far more complex (see Farrar [2]). At a basic level, the control problem with flexible workers can be regarded as a problem of service rate control, with service rate proportional to the fraction of worker effort applied at a given station. Work in a similar spirit can be found in Kim [3], in which tandem queueing systems and systems with job forking are addressed in the context of flexible workers. Bartholdi and Eisenstein [1] consider a deterministic version of our model with the additional restriction that only one worker may occupy a station at any given time. Zavadlav et al. [9] show that a flexible assignment system can outperform fixed assignments in a variety of similar circumstances, using Markovian and simulation models. Mandelbaum and Reiman [5] have taken a performance analysis approach to the assessment of queueing systems in which workers are pooled. Their work treats cycle time evaluation in the class of Jackson networks, comparison of pooled server systems with specialized/static models, and parallel tasks.

Traditional "rigid" models that include scheduling, routing, or admission control typically use knowledge of the distribution of work in the system to make decisions regarding which job to perform next. We will use the term "flexible worker" to describe models in which the human or automated worker can (1) perform work at more than one workstation which may require multiple skills, or (2) use knowledge of the distribution of work in the system to make decisions regarding which job to perform next, or (3) work in cooperation with other workers to increase the production rate of all of the possible jobs.

We believe that the impact of flexible workers is already being widely seen in industry. For example, at John Deere Corporation, a heavy equipment manufacturer, production work-cells have been created in which teams of workers cooperate in a sequence of steps that result in the manufacture and assembly of a subsystem. To enable this approach to

---

[1] This material is based upon work supported by the National Science Foundation under Grant No. DMI-9522795 to Northwestern University.

production, each of the workers have been empowered in the following ways: (1) he/she has been trained for the full array of skills and operations required at any stage in the process, (2) sufficient equipment investment has been made so that he/she has access to the resources needed for any job without blocking or being blocked by another worker, (3) he/she is paid a relatively high wage, as dictated by the wage scale for the highest paid skill required in the workcell, and (4) he/she is expected to actively participate in determining appropriate and effective ways to accomplish the production process. It stands to reason that workers who are "cross-trained" to perform various types of work will frequently require higher labor rates on average than in a system with a rigid division of labor. Nevertheless, we will provide evidence that the significant performance gains may outweigh the cost.

Provided analysis can be performed for idealistic systems such as this, the results will provide fundamental bounds on the achievable performance. Such results are useful in understanding what classes of systems offer little potential improvement using cross-training and flexible workers. Secondly, they also provide a context in which researchers and practitioners can develop effective and practicable approaches that raise performance above that of past approaches and closer to the performance bound. Another good reason was given in by Solberg [6]: "By recognizing that the power of a model is directly related to its simplicity (rather than its complexity), we may be able to redirect some of our attention to new formulations."

The objectives of this paper are to formulate an appropriate model to identify some of the fundamental limits of performance achievable with cross-trained workers, and to suggest approaches for obtaining effective worker allocation with policies which can be easily implemented.

2. Problem Formulation

We consider a series network of queues; that is, jobs served at queue n proceed directly to queue $n+1$ and from station N they exit the system. We allow a general (possibly time-varying and correlated) arrival process under the following assumption. The cumulative number of arrivals up to and including time t, denoted by $A(t)$, is such that $A(t) \in \mathbb{N}$ for any $t \in \mathbb{R}^+$, where $\mathbb{N}(\mathbb{R}^+)$ denotes the naturals (nonnegative reals). Thus, a finite system queue length is guaranteed for any finite time. Let $a(i)$ denote the random time of the arrival of the ith job, where we require $E\{a(i)\} < \infty$ for all $i \in \mathbb{N}$. Batch arrivals are permitted, but we assume that they all have distinct labels that dictate the arrival sequence. This is important, because we assume that job completions must be first come, first served (FCFS) at the last station only.

We make the following idealistic assumptions about workers. All workers can cooperate on the same job without interfering with each other. At any time t, the total worker effort applied in the system cannot exceed the maximum amount of flexible effort available: M. We assume that the workforce is divisible among the N stations in arbitrary increments; therefore, we allow $M \in \mathbb{R}^+$. We make the important assumption that workers can move from one station to another with zero cost and zero time (as was done in the analysis of [1], [5], and [3]). We model a process of underlying *service requirements* that are independent of all else in the system. At station n the service requirements for successive jobs $i = 1, 2, 3, \ldots$ are given by the i.i.d. sequence of random variables $\{S_n(i) : i \in \mathbb{N}\}$ with the unit being man-hours. The cumulative distribution function (c.d.f.) of $S_n(\cdot)$, denoted by F_n, is a general strictly positive distribution with finite first and second moments. We let $\mu_n = (ES_n(i))^{-1}$ denote the mean rate of the underlying service requirement at station n. We assume that work effort at any station or job can be controlled continuously over time and that the time it takes to complete a job at any station is inversely proportional to the amount of effort allocated. Due to the generality of the arrival process and the service time distribution, our claims to optimality must be justified within the context of G, the class of all nonanticipative policies based on the history of observed states and control actions through time t. This requires a highly detailed state space. The system state is denoted by the matrix $(q(t), p(t))^T$, where T indicates transpose. Let $q(t) = (q_1(t), q_2(t), q_3(t), \ldots)$ with $q_i(t) \in \{0, 1, 2, \ldots, N, N+1\}$ denoting the queue that job i is in at time t, 0 signals that the job has not yet arrived, and $N+1$ signals job completion. Let $p(t) = (p_1(t), p_2(t), p_3(t), \ldots)$ with $p_i(t) \in I\!R^+$ denoting the amount of the underlying service requirement at station $q_i(t)$ that has been met on $[0, t]$. From this state, one can generate $x(t) \triangleq (x_1(t), x_2(t), \ldots x_N(t))$, the state vector of queue lengths (including any job in service) at time t via $x_n(t) = \sum_{i=1}^{\infty} 1\!\!1\{q_i(t) = n\}$ with $1\!\!1\{\cdot\}$ denoting the indicator function.

A Markov control policy g maps $(q(t), p(t))^T$, to an effort allocation matrix $u(t) \triangleq [u_{n,i}(t)] : n = 1, 2, \ldots, N$ and $i = 1, 2, 3, \ldots$. We require $u(t) \in \mathcal{C}$, the space of effort matrices with $\sum_{n=1}^{N} \sum_{i=1}^{\infty} u_{n,i}(t) = M, u_{n,i}(t) \geq 0 \ \forall i$. If at time t, $q_i(t) = n$ and $p_i(t) = 0$, then job i will be completed at station n at the random time $C_n^g(i) = \inf\{T \geq 0 : \int_t^T u_{n,i}^g(t)dt = S_n(i)\}$. If in addition we assume that $u_{n,i}^g(t) = \underline{u} \in I\!R^+$ for a time interval of at least $S_n(i)/\underline{u}$ along each sample path, the resulting mean of the *service time*, $\tilde{S}_n^g(i)$, at station n is

$$E\{\tilde{S}_n^g(i)\} = E\{S_n(i)\}/u_n = \frac{1}{\underline{u}\mu_n}. \tag{2.1}$$

An outcome of this model is that the coefficient of variation (cv) for the service time with a fixed effort level u_n at station n is $(\sqrt{var(S_n)/\underline{u}})/(E\{S_n\}/\underline{u})$. This implies that the cv is unaffected by the effort allocation, which is intuitive.

We proceed to investigate optimal server allocation policies under the assumption that perfect information of the queue length vector is provided to a centralized controller, which executes the policy. As we will see, an optimal policy is sufficiently simple that it can be implemented even by independent agents.

3. Optimality of the Expedite Policy

Having defined the problem, we proceed to demonstrate that a simple policy is optimal with respect to job completion times along every sample path. Let the random variable $C_N^g(i)$ denote the completion time of the ith job at the last station. Specifically, all of the effort should be allocated to the same job so as to bring the current job from queue 1 through its completion at queue N without interruption. We call this policy the *expedite policy*, denote it by XP, and define it as $u_{n,i}(t) = M$ if-f $q_i(t) = n$ and $q_{i-1}(t) = N+1$. Note that if $M \in I\!N$, XP assigns all workers to a job and they proceed to bring jobs through the system one at a time. Thus, there is no need to make use of the idealistic assumption that workers are "infinitely divisible".

Theorem 1: Assuming that job completions must obey FCFS discipline at station N, the expedite policy achieves minimum completion times for all jobs along every sample path:

$$C_N^{XP}(i) \leq C_N^g(i) \quad \text{a.s.} \quad \forall i \in I\!N \ .$$

Proof: We use sample path coupling to show that for any realization of the random variables, ω, $C_N^{XP}(i, \omega) \leq C_N^g(i, \omega)$, for all $g \in G$, and $i \in I\!N$. Policy $g \in G$ is a *nonidling policy* if $\sum_{i=1}^{\infty} u_{n,i}^g(t) = 0$ whenever $x_n(t) = 0$ and if $x_i(t) > 0$ for some node i and time t, then all M workers must be allocated to the nonempty stations. We will only show the result for "nonidling policies", since the performances of all other admissible policies are dominated by those of nonidling policies. It is apparent that the expedite policy we defined above is a nonidling policy. Assume the first job arrives at time 0. Under a realization $\omega \in \Omega$ expressed in terms of arrival and service requirement times, we have $a(i, \omega)$ denoting the arrival time of the ith job, $S_n(i, \omega)$ denoting the service requirement time of the ith job at station n, and let $C_N^g(k,\omega)$ be the time that the kth job exits the system. Under policy XP, the completion time of the kth job will be

$$C_N^{XP}(k,\omega) = max\{C_N^{XP}(k-1,\omega), a(k,\omega)\} + \sum_{n=1}^{N} \frac{S_n(k,\omega)}{M} . \tag{3.1}$$

For job 1, the claim is obvious. Since XP policy allocates all of the effort towards the processing of job 1, it achieves the minimum achievable completion time for the job,

$$C_N^{XP}(1,\omega) = \sum_{n=1}^{N} \frac{S_n(1,\omega)}{M} . \tag{3.2}$$

As the induction hypothesis, assume that XP achieves the minimum achievable completion times for jobs $1, 2, \ldots, k-1$. If job k arrives to an empty system, the argument for job 1 applies to job k. If job k arrives during a busy period to a system with l jobs, $l \in \{1, 2, \ldots, k-1\}$ ahead of it, then due to the FCFS discipline assumption, we know that all l jobs must be completed before job k.

Now, it is obvious that idling for a positive length of time is suboptimal and that no effort should be allocated to jobs $\{k+1, k+2, \ldots\}$ (stealing from jobs $\{k, k-1, \ldots, k-l\}$), prior to the completion of job k, since this is only going to delay the completion of jobs $k, k-1, \ldots, k-l$.

Let $g \in G$ be any nonidling policy with FCFS discipline at station N, which allocates some effort to job k prior to the completion of job $k-r$, where $r \in \{1, 2, \ldots, l\}$. To construct an interchange argument, we can exchange equal amounts of work between jobs $k-r$ and k as follows: Let $C_N^g(k-r,\omega)$ be the completion time of job $k-r$ under realization $\omega \in \Omega$ on probability space $(\Omega, \mathcal{S}, P)$. Then, there must be at least an $\epsilon_1 > 0$ amount of effort allocated to job $k-r$ over an interval of some length $\epsilon_2 > 0$ prior to its completion at $C_N^g(k-r,\omega)$. Similarly, by assumption, g must devote at least an $\epsilon_3 > 0$ amount of effort to job k over some time interval of length $\epsilon_4 > 0$ prior to $C_N^g(k-r,\omega)$, which we will designate $(t, t+\epsilon_4)$. Let $\epsilon = min\{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$. To construct a modified policy g', we take ϵ amount of effort from job k during the interval $(t, t+\epsilon)$ and allocate this ϵ^2 man-hrs towards the processing of job $k-r$ during the same time period. Similarly, we take ϵ amount of effort from job $k-r$ and allocate it to job k during the interval, $(C_N^g(k-r,\omega)-\epsilon, C_N^g(k-r,\omega))$. Due to the assumption of FCFS completions at station N, job k must be available for processing during $(C_N^g(k-r,\omega)-\epsilon, C_N^g(k-r,\omega))$.

It is easily seen that under g', the completion time of job $k-r$ will decrease by Δ, where Δ can be computed from the equation

$$\epsilon^2 = \int_{C_N^g(k-r,\omega)-\Delta}^{C_N^g(k-r,\omega)} [\sum_{n=1}^{N} u_{n,(k-r)}^g(t)]dt . \tag{3.3}$$

We observe that the completion time of job k does not change under policy g', since the interchange applied only "shifted" the work done on job k from $(t, t+\epsilon)$ to $(C_N^g(k-r,\omega)-\epsilon, C_N^g(k-r,\omega))$, which did not change the amount of processing done on job k prior to $C_N^g(k-r,\omega)$.

Thus, we conclude that allocating effort to job k prior to the completion of the l jobs ahead of it only results in delaying any or all of the jobs $\{k-1, k-2, \ldots, k-l\}$ and does not change the completion time of job k.

Repeating the interchanges iteratively, we reach the XP policy, and due to the result above, we conclude that XP must achieve the minimum completion times for jobs $\{1, 2, \ldots, k-1, k\}$ and thus, the claim is proven inductively along every sample path in probability space $(\Omega, \mathcal{S}, P)$. □

Theorem 1 has implications for the throughput, work-in-process, and cycle time performances of the system.

Corollary 1: In the almost sure sense, the XP policy maximizes the throughput $\Theta^g(t) \triangleq max\{k \in I\!N : C_N^g(k) \leq t\}$, minimizes the WIP $L^g(t) \triangleq |x^g(t)| = \sum_{n=1}^N x_n^g(t)$ for $t > 0$, and minimizes the cycle times $W^g(i) \triangleq C_N^g(i) - a(i)$, $i \in I\!N$.
Proof: First we observe that since $C_N^{XP}(i) \leq C_N^g(i)$ a.s., it is immediate that
$$max\{i \in I\!N : C_N^{XP}(i) \leq T\} \ \geq \ max\{i \in I\!N : C_N^g(i) \leq T\} \text{ a.s.}$$

Next, defining the WIP level as $\sum_{n=1}^N x_n^g(t) = A(t) - \Theta^g(t)$, from the preceding case we see that the WIP level is also minimized.

Because $W^g(i) = C_N^g(i) - a(i)$ and the arrival time, $a(i)$, is uncontrollable for all jobs $i \in I\!N$, the result for cycle times of jobs is immediate from Theorem 1. □

Independent of our research, Kim [3], addressed the problem described above and argued that the XP policy is optimal. Unfortunately, there are some difficulties with the details of her proof, which was never published. In addition to the result for service rates that are linear in the number of servers, Kim makes the important point that if the service rate of m workers at station n is convex in m, then the XP policy remains optimal. Unfortunately, such is not the case with service rates that are concave in the number of workers, as the following example illustrates:
Example: Consider a deterministic system in which one job arrives at each instant ($a(k) = k, k \in I\!N$) to a two-station system with two workers. At each station, the service times have the same concave structure: if one worker is present, then the service rate is 1; if two workers are present, then the service rate is $1+ \epsilon'$ such that the resulting service time is $1 - \epsilon$ for some small $\epsilon < 1/2$; and otherwise the service rate is zero. At time zero there are no jobs in either station.

If one considers a fixed allocation with one worker fixed at each station, then the following excellent long run performance results: the throughput of the system is 1 job per unit time, the cycle time is 2 units, and the average WIP is 2 jobs (one at each station).

On the other hand, the XP policy assigns both workers to the most recent arrival. Thus the completion time of the kth job occurs at time $C_N^{XP}(k) = k*2*(1-\epsilon)$. Thus the advantage of the static/fixed allocation policy becomes $C_N^{XP}(k) - C_N^s(k) = k*2*(1-\epsilon) - k$, which goes to infinity as k goes to infinity. □

It is interesting to observe that the XP policy remains optimal under the assumption that jobs are never rejected if the system buffers are finite for some of the queues. Observe that XP is a policy that maintains jobs in queue (and not in service) only at station 1, and at any instant keeps only one job in service. By Corollary 1, XP minimizes the total WIP, $L(t)$, (queue plus service) pathwise. Thus, the XP policy is optimally efficient from the standpoint of buffer design. It requires a buffer only at station 1 and it uses less total buffer space across stations than any other policy. Thus we have the following property:
Corollary 2: For a given problem instance with flexible workers and without the occurrence of job rejections pathwise, the policy in G that minimizes (pathwise) the total buffer space required is XP; moreover, all of the buffer space required by XP can be pooled at station 1.

For most open queueing networks, the pathwise absence of rejections is an extremely strong assumption that will not hold up. On the other hand, it applies to deterministic (time-varying) systems and to some stochastic systems with bounded service times and strict lower bounds on the interarrival times.

In many applications, it is realistic to assume that the worker staffing level will vary over time due to absenteeism, vacations, maternity leaves, up/down-sizing, changes in demand, and so forth. Let us extend the model from a constant number of workers M, to a total staff level $M(t)$ that varies as a function of time. We allow the staff level process

$\{M(t) : t \in \mathbb{R}^+\}$ to be either deterministic or stochastic, provided that the sample paths are continuous almost everywhere. For the sake of generality, we also allow the service times (underlying service requirements) to depend on the job index i, station number n, and past service completions of other jobs at other stations. It is important that the effect of multiple workers on the combined service rate be linear.

Under this increased generality, the proof of the optimality of XP remains valid as given above. The XP policy processes the farthest downstream job with maximum possible service rate, therefore maintaining the following result:

Corollary 3: The optimality claims made in Corollary 1 for XP remain true with a time-varying staff level and general service processes.

The XP policy is an easily implemented self-organizing policy in the sense that it does not require any computation for allocation of servers to stations and jobs. Management can adjust the production rate by changing the staffing levels without disturbing the system's operation. In addition it results in an equal utilization of all workers and gives a nonidling policy. Unfortunately, many systems do not allow multiple workers to process the same job, and we anticipate that significant further research is needed to understand the resulting problems.

4. Comparison of the Optimal Flexible Policy with the Optimal Static Allocation

We are particularly interested in addressing the following question. Since most systems are designed with a fixed/static allocation of workers/capacity, what performance increase can be gained through the use of flexible workers/capacity without increasing the total effort available?

We restrict attention to a Poisson arrival process with rate λ and exponential(μ_i) service times at each queue. As before, the total effort allocation capacity is M. To assess the benefits of flexible workers, we compare the optimal cycle time achievable with a fixed worker allocation to that achievable with flexible workers. Under the exponential processing time assumption, the network is a special case of a Jackson network. In steady state, the network behaves as N tandem independent M/M/1 queues (see Wolff [8]). The mean cycle time for a static allocation $u \in \mathbb{R}^N$, $u_n \geq 0$ is given by

$$W^s(u) = \sum_{n=1}^{N} (u_n \mu_n - \lambda)^{-1}. \tag{4.1}$$

The effective service time of the ith job under XP is given by the random variable $\sum_{n=1}^{N} S_n(i)/M$. The mean cycle time of the XP policy can be computed as in Wolff [8] using the Pollaczek-Khintchine formula:

$$W^{XP} = \frac{\lambda[var(T_{XP}) + (E[T_{XP}])^2]}{2(1 - \lambda E[T_{XP}])} + E[T_{XP}]. \tag{4.2}$$

We can make this dramatic simplification because the system boils down to an M/G/1 queue with mean service time, $E[T_{XP}] = \sum_{n=1}^{N} (M\mu_n)^{-1}$ and variance, $var(T_{XP}) = \sum_{n=1}^{N} (M\mu_n)^{-2}$ under the assumption of independent of service requirement times at stations.

Although a direct comparison can be made for any static allocation, we optimize the static allocation under the constraint that $\sum_{n=1}^{N} u_n = M$. The same result can be found using Kleinrock's square-root channel capacity assignment formula (see section 5.7 of Kleinrock [4]), as has been noted in Mandelbaum and Reiman [5]. We present the optimization model here for the sake of completeness.

Theorem 2: Assume that the rate of the Poisson arrival process, $\lambda \in I\!R^+$ is such that $1 - \lambda \sum_{i=1}^{N} (M\mu_i)^{-1} > 0$. Then, the optimal static allocation of workers that minimizes the average total cycle time is given as

$$u_n^* = \frac{M - \lambda \sum_{i=1}^{N} \mu_i{}^{-1}}{\sqrt{\mu_n} \sum_{i=1}^{N} \mu_i{}^{-1/2}} + \frac{\lambda}{\mu_n}, \qquad n = 1, 2, 3, \ldots, N. \tag{4.3}$$

Proof: The problem can be written as follows:

$$\begin{array}{ll} \min & W^S = \sum_{i=1}^{N} (u_i \mu_i - \lambda)^{-1} \\ \text{subject to} & \sum_{i=1}^{N} u_i = M \\ & u_n > 0, \quad n = 1, 2, \ldots, N \end{array} \tag{4.4}$$

We relax the positivity constraint to a nonnegativity constraint and solve the resulting optimization problem using the method of Lagrange multipliers.

Using the assumption that $1 - \lambda \sum_{i=1}^{N} (M\mu_i)^{-1} > 0$, it is trivial to show that the vector $u^* = [u_1^*, u_2^*, \ldots, u_N^*]'$ as given above satisfies the first order necessary and sufficient conditions and the Hessian of the Lagrangian is positive definite. Thus, u^* is globally optimal for the relaxed problem. Since $1 - \lambda \sum_{i=1}^{N} (M\mu_i)^{-1} > 0$ by assumption, $u_n^* > 0$ for all $n = 1, 2, \ldots, N$, u^* is also feasible for the original problem, and thus, it is the globally optimal static allocation that minimizes the average total cycle time. □

From the previous result, the best mean cycle time that a fixed allocation system can achieve is:

$$W^{S^*} = \frac{(\sum_{n=1}^{N} (\mu_n)^{-\frac{1}{2}})^2}{M(1 - \lambda E[T_{XP}])} \tag{4.5}$$

where $E[T_{XP}] = \sum_{i=1}^{N} (M\mu_i)^{-1}$ by definition.

The stability boundary, ($\rho = 1$) is captured for both the expedite policy and the optimized static allocation by the condition $\lambda < E[T_{XP}]$. For a given vector of mean underlying service requirement rates, μ, the mean cycle time performances of both flexible and static allocation systems do not change as the order of the μ_i coefficients are permuted. We refer to the resulting difference between the optimal static allocation cycle time (W^{S^*}) and the cycle time of the XP policy (W^{XP}) as *opportunity*, which is given by:

$$W^{S^*} - W^{XP} = \sum_{1 \le i < j \le N} \frac{2(\mu_i \mu_j)^{-\frac{1}{2}} + \lambda (M\mu_i\mu_j)^{-1}}{M(1 - \lambda E[T_{XP}])} \tag{4.6}$$

This measure of opportunity is conservative, because we are accounting for the minimum cycle time achievable with a static allocation. Nevertheless, stability guarantees that the denominator is positive, thus, the difference is always positive, verifying that a flexible worker system outperforms a rigid one under the conditions assumed for XP.

Having quantified opportunity, we wish to address the question of which conditions maximize the opportunity, since the answer will give us some insights about systems where flexible workers would result in better system performance.

Definition: A balanced system is one with $\mu_i = \mu_j$ for all $i, j = 1, 2, \ldots, N$.

Theorem 3: For an N-station tandem queueing system, a balanced system maximizes the difference between the average total cycle time of an optimal static allocation and that of a flexible system under the expedite policy.

Proof: Using the expression for the opportunity derived above, the problem can be formulated as follows (P1):

$$\begin{array}{ll} \max & \sum_{1 \le i < j \le N} 2(\mu_i \mu_j)^{-\frac{1}{2}} + \lambda (M\mu_i\mu_j)^{-1} \\ \text{subject to} & \sum_{i=1}^{N} (\mu_i)^{-1} = M E[T_{XP}] \\ & \mu_n \ge 0, \qquad n = 1, 2, \ldots, N. \end{array} \tag{4.7}$$

This formulation relaxes the positivity constraint (mean service rates must be strictly positive at all stations) to a nonnegativity constraint, as in the previous problem. As can be seen from the model, as $\mu_i \to 0$ for any i, $\exists\ j$ such that $1/\mu_j \to -\infty$. For that reason, the model will never assign 0 to any of the μ_i, and thus, the above formulation works.

The theorem states that a "balanced system" maximizes the difference, or the opportunity. In other words, the claim is that $\mu^* = (\frac{N}{ME[T_{XP}]}, \frac{N}{ME[T_{XP}]}, \cdots, \frac{N}{ME[T_{XP}]})$.

To solve the model, we relax the second constraint of (P1) and check if the solution to the relaxed model is also feasible for the original model. We eliminate the first constraint by writing μ_N in terms of the other decision variables as $\mu_N = (ME[T_{XP}] - \sum_{i=1}^{N-1} \mu_i^{-1})^{-1}$.

It is straightforward to see that μ^* satisfies the first order necessary conditions and the Hessian is negative definite when evaluated at μ^*. Since the vector is also strictly positive, it is feasible for the original problem. Thus, we conclude that μ^* uniquely maximizes the opportunity. □

From a managerial perspective, it is interesting to identify conditions under which opportunity is increased. Although opportunity is always computed with respect to optimized worker allocations, systems in which the underlying service requirements are unbalanced will provide less opportunity. To give an example, we considered a problem with 5 stations and 6 servers. We compared a "balanced" system with an "unbalanced" system. Although we do not have an accurate measure of *system unbalance*, computational experience suggests that the following index B is roughly effective in capturing what drives the relative value of opportunity:

$$B = \frac{\sum_{n=1}^{N} \mu_n - N min_n\{\mu_n\}}{\sum_{n=1}^{N} \mu_n} . \tag{4.8}$$

The index B is zero for a perfectly balanced system, and approaches one in the limiting case of unbalance. For the balanced case, we selected underlying service rate requirements $\mu_n = 8.0$ for all n (which specifies complete balance). With six servers and five stations, the system will be stable provided that the arrival rate is such that $\lambda \leq 48/5$. To provide a sense of the range of possibilities, we also present an extreme case of unbalance. To provide a fair comparison, we selected parameters which would provide stability over the same region ($\lambda \leq 48/5$). Thus, we selected $\mu_1 = 2650$, $\mu_2 = 1.6$, $\mu_3 = 5320$, $\mu_4 = 3580$, $\mu_5 = 3580$, which is clearly highly unbalanced. For each case, we computed (exactly) the difference $W^{S^*} - W^{XP}$ and we present the results in Figure 1. The graph illustrates the fact that opportunity increases with system utilization and also as the underlying service requirements become increasingly balanced.

5. Analysis of Closed Networks

Having analyzed the performance of open queueing networks in the previous sections, it seems appropriate to examine closed queueing network (CQN) models. We may suppose that there are K jobs in the network, with jobs served at station N being sent directly back to the queue of station 1.

Closed networks arise naturally in systems for which there is unlimited availability of raw materials and an admission controller which releases a new job to the system whenever a job completes its processing, thereby maintaining a CONstant Work In Process (CONWIP) release of work to the system with K jobs at all times.

We formulate the CQN model as follows. There are N identical stations, each having an exponentially distributed underlying service requirement with rate 1. We assume that the total worker staff is of constant size M. We contrast the XP policy with a static policy, S. If M is not a multiple of N, the static system is assumed to be able to split workers (a generous assumption) so that the number of workers assigned to any station is exactly M/N. We assume for both policies that if m workers are present at a station,

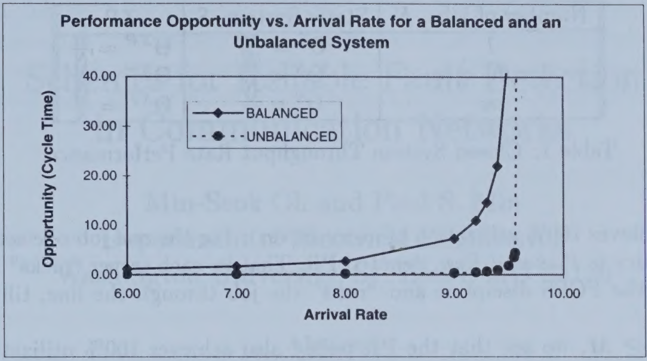


Figure 1: Opportunity, $W^{S^*} - W^{XP}$, versus λ for a balanced and an unbalanced system.

the resulting mean service time is $1/m$. The throughput rate of a policy is taken to be the long run rate of job completions at station N.

The analysis of the XP policy is straightforward, since there is no worker starvation. Every $1/M$ units of time on average, one job (the one farthest downstream) is advanced one station. Thus, one job is completed every N/M units of time. It is important to observe that the throughput rate of XP is insensitive to the number of jobs, K: $\Theta^{XP}(K) = \Theta^{XP}(1) \ \forall \ K \in \mathbb{N}$.

The analysis of the static system, S is straightforward for $K = 1$, since there is no queueing of jobs. Every 1 unit of time on average, the job is advanced one station. Thus, a job is completed every N/M units of time at each station, resulting in a cycle time of N^2/M. It is important to observe that the throughput rate of S is sensitive to the number of jobs, K.

In the case where $K = \infty$, we see that every buffer of the static system will always have jobs present. Therefore, every $(M/N)^{-1}$ time units, a job exits station N, on average. The throughput rate becomes M/N. With $K < \infty$, worker starvation will occur with empty queues and M/N can serve as an upper bound on throughput rate.

These performance measures are summarized in Table 1. We note that the XP system achieves with $K = 1$ a throughput rate that is achievable in the static system only with $K = \infty$. On the other hand, as system congestion increases (heavy traffic), the XP policy's performance does not improve and is approached by the performance of static systems.

When the number of jobs is not less than the number of workers ($K \geq M$), it is interesting that a sensible policy other than XP can achieve the same performance in a CQN. Before introducing this new policy, let us provide its motivation. Many applications may restrict the ability of multiple servers to simultaneously serve the same job. Suppose, for the sake of argument, that only one server can attend any job at any time. We retain the assumption that machines (tools, jigs, floorspace, etc.) are in sufficient supply that all workers can work simultaneously on distinct jobs at the same station (and thus, there is no blocking of one worker by another). We also assume workers are completely flexible (can work at any station) and all have the same pace. That is, for each station n, the work rate is the same for each server and is taken to be 1 unit without loss of generality.

The expedite policy has been proven optimal for open networks under the assumption that all the servers can attend any job simultaneously (this is clear for CQN's since the

Number of jobs, K	Static System, S	XP
1	$\Theta^S = \frac{M}{N^2}$	$\Theta^{XP} = \frac{M}{N}$
$1 < K < \infty$	$\Theta^S < \frac{M}{N}$	$\Theta^{XP} = \frac{M}{N}$
∞	$\Theta^S = \frac{M}{N}$	$\Theta^{XP} = \frac{M}{N}$

Table 1: Closed System Throughput Rate Performance.

XP policy achieves 100% utilization of every server). For the one-job-one-server case, an analogous policy is *Pick-and-Run*, denoted PR. That is, each server "picks" the next job according to the FCFS discipline and "runs" the job through the line, till it exits the system.

With $K \geq M$, we see that the PR policy also achieves 100% utilization of every server. The throughput rate of any given server is $1/N$. Since all M workers serve in parallel without blocking, the throughput rate is $\Theta^{PR} = M/N = \Theta^{XP}$ (and the cycle time remains at KN/M by Little's Law).

6. Conclusion

In this paper, we provided a careful formulation of an idealistic model for flexible workers in queueing networks. For series queueing networks with general assumptions, we proved that a simple policy, the expedite policy, optimizes along each sample path the cycle time for each job. Therefore, the expedite policy also minimizes work in process and maximizes throughput rate along every sample path for any time t. We considered the optimal static allocation of effort and used this to compute the performance improvement achievable using flexible workers, and we demonstrated that equal service rates maximize the opportunity of flexible workers. Closed queueing network models were used as another perspective that provides an assessment of the benefits of flexible workers.

References

[1] J.J. Bartholdi, III and D. D. Eisenstein, A production line that balances itself, *Operations Research*, **44:1** (1996) 21–34.

[2] T.M. Farrar, Optimal Use of an Extra Server in a Two Station Tandem Queueing Network, *IEEE Transactions on Automatic Control* **AC-38**, (1993) 1296–1299.

[3] J.M. Kim (1993) *Optimal Design and Control of Queueing Networks*, Ph.D. Thesis, Rutgers, The State University of New Jersey.

[4] L. Kleinrock (1976) *Queueing Systems, Volume II: Computer Applications*, Wiley, New York.

[5] A. Mandelbaum and M.I. Reiman, On Pooling in Queueing Networks, To appear in *Management Science*, (1997).

[6] J.J. Solberg, The power of simple models in manufacturing, in *Manufacturing Systems: Foundations of World-class Practice*, J. A. Heim and W. D. Compton, ed's. (National Academy Press, Washington, D.C. 1992).

[7] R.W. Wolff, *Stochastic Modeling and The Theory of Queues* (Prentice-Hall, Englewood Cliffs, 1989).

[8] E. Zavadlav, J.O. McClain and L.J. Thomas, Self-buffering, self-balancing, self-flushing production lines, *Management Science*, **42:8** (1996) 1151–1164

Schemes for Reliable Fault Detection in Communication Networks

Min-Seok Oh and Paul S. Min
Department of Electrical Engineering
Washington University, St. Louis, MO 63130

Abstract

We investigate how to construct a reliable alarm structure in communication networks. First we study how to allocate a minimal set of alarm collecting devices and still cover all critical alarms in mesh networks under the assumption that alarms are delivered along provisioned paths. We also consider probabilistic nature of alarm loss. For a high level of reliability in communication, multiple faults are considered since there is high probability for alarms to be lost. We investigate how to provision paths in mesh network while reducing alarm loss under the multiple fault condition.

1 Introduction

In communication networks it is hoped that all generated alarms are reported to fault management centers without being lost. The possibility of losing alarms in the network is one of the main difficulties to overcome in network management. There are two cases where an alarm emitted by a component is not received by a supervising system; *masking* and *saturation* [2]. Each alarm must be delivered to a network manager even though the manager may not use it for localization process. When a network manager has insufficient alarms, fault localization process will take longer and output inaccurate hypotheses for the cause of the alarms. One of the efforts to improve the alarm delivery rate will be to allocate alarm collecting devices (ACDs) properly so that they can report a failure in a network component to the fault manager. We propose a scheme which allocates a minimal set of ACDs and still covers all critical alarms in mesh networks under the assumption that alarms are delivered along provisioned paths.

In a general network topology, an alarm loss rate is closely related to the hop distance from the ACDs since remote alarms experience more masking and saturation than nearby alarms. We can keep the alarm loss rate below a certain level by deploying more ACDs for a high level of reliability. Multiple link failures on shared paths can prevent the alarms caused by one of the link failure from being delivered to the ACD, if a link failure close to the ACD blocks the alarm from the distant link failure. ACDs should be allocated so that any critical link failure in the network can be reported to at least one of the ACDs. Although ideally ACDs can be placed at every node, it is not preferable since a large number of ACDs increase network management costs and the traffic between ACDs and a management center. Hence as a second effort for reliable communication, we propose a scheme which provisions paths in mesh network while reducing alarm loss under multiple fault condition.

There are many types of faults. To make various issues in the fault management clear, in this paper, we will mainly focus on link failure in a network. Link failures may not only be caused by breaking down of links physically, but also by faults occurring at the link-level protocol. In [1] and [5], it is reported that the loss of system connectivity, i.e., link failure, accounts for a very significant fraction of the network management problems.

2 Allocation of ACDs

For reliable communication, all the alarms generated by a faulty component should be reported to the fault management center. Therefore it is essential to allocate ACDs at appropriate locations in order to cover all critical alarms in a network.

We assume that when there is a link failure, the alarms are reported to all the nodes which are passed by the virtual path on the failed link. As a real example, in SONET *Alarm Indication Signal* (AIS) is used to alert downstream nodes and *Remote Defect Indication* (RDI) alerts the upstream nodes for a link failure [6].

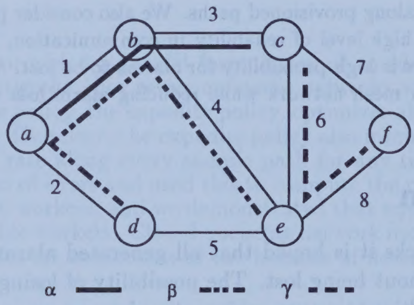


Figure 1: An example of a mesh network.

In order to describe the network topology and provisioned paths symbolically, we use the incidence matrix and introduce the provisioning matrix.

The incidence matrix C of a graph G is formed in the following way. A column of the matrix is allocated to each arc and a row to each node. An entry at (i,j) is 1 if node i is an endpoint of arc j and is zero otherwise. The incidence matrix for the graph of Figure 1 is

$$\mathbf{C} = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left[\begin{array}{cccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right] \end{array} \quad (1) \qquad \mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccc} \alpha & \beta & \gamma \end{array} \\ \left[\begin{array}{ccc} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{array}\right] \end{array} \quad (2)$$

In the matrix $\mathbf{P}_{l \times p}$, where l represents the number of links in the network and p the number of paths provisioned, the entry at (i,j) is 1 if path j uses link i, 0 if not. For a network, where three paths are provisioned, as shown in Figure 1, the provisioning matrix $\mathbf{P}_{(l \times p)}$ can be written as (2).

The product $\mathbf{PP}^{\mathrm{T}}$ gives a symmetrical matrix. Each element at (i, j) represents the total number of paths on link i which is passed through by the paths using link j.

If we multiply $\mathbf{C}$ with $\mathbf{PP}^{\mathrm{T}}$, the resulting matrix becomes a matrix which indicates reachability between nodes and links.

$$\mathbf{CPP}^{\mathrm{T}} = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left[\begin{array}{cccccccc} 3 & 2 & 1 & 2 & 0 & 0 & 0 & 0 \\ 4 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 2 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{array} \right] \end{array} \tag{3}$$

In other words, when there is a link failure at a link, the failure on the corresponding column can be detected by the nodes on non-zero rows in $\mathbf{CPP}^{\mathrm{T}}$. For example, when link 4 has failed, we see from (3) that it can be detected at node a, b, d and e. The analysis is as follows: Finding all the links which can be reached by a node through provisioned paths is equivalent to finding all the links which can be reached by the links connected to the node because a path always ends at a node. A column j in $\mathbf{PP}^{\mathrm{T}}$ provides information on the reachability from a link j to other links. And a row i in the incidence matrix $\mathbf{C}$ gives information on links connected to a node i. Therefore the product $\mathbf{CPP}^{\mathrm{T}}$ provides information on reachability from a node to a link in a network.

By using the matrix $\mathbf{CPP}^{\mathrm{T}}$ we can allocate ACDs which cover all single link failures. Note that in this particular case a link failure at link 5 and link 7 will never be detected, since all elements are zero in the 5th and 7th columns, which means that no path passes link 5 and link 7. Therefore link failures on those links will not be reported to the management center with the current alarm reporting scheme. Since the failures in those links do not affect the normal network operation, they are called *non-critical* faults.

As the network grows, the number of ACDs is subject to increase. For the sake of low network management cost and traffic, it is appropriate to use a minimal number of ACDs, which cover all critical link failures. This can be achieved by using a coverage chart similar to the prime implicant chart which is used in simplifying and minimizing the boolean functions [4].

We can construct a link coverage chart as shown in Table 1 by replacing non-zero elements in $\mathbf{CPP}^{\mathrm{T}}$ with an × mark. We can apply the same methods used in simplifying the boolean function to find the minimum number of nodes covering all the critical links [3].

Table 1: Link coverage chart.

	1	2	3	4	6	8
a	×	×	×	×		
b	×	×	×	×		
c	×		×		×	×
d	×	×		×		
e	×	×		×	×	×
f					×	×

3 Allocation of ACDs given Alarm Loss Probability

As explained in Section 1, alarms are lost for two reasons, masking and saturation. Once probabilities for alarm loss can be estimated, it is possible to keep the loss probabilities below a certain level for reliable communication by reallocating ACDs or adding more.

We impose following assumptions.

1. The alarm loss probability is closely related to the hop distance from an ACD which is responsible for the failure at a particular link.
2. When there are multiple paths on a link the event that an alarm for a particular path on the link is lost is independent of the event that an alarm for the other path on the link is lost. This is true if the paths do not share any other link and failures of links take place independently. Hence, if the alarm loss probabilities for k paths at a link seen by ACDs are $p_1, p_2, \cdots, p_k$ respectively, the combined alarm loss probability will be the product of them, i.e., $p_1 p_2 \cdots p_k$.

We can construct a table similar to the link coverage chart earlier, but we write the alarm loss probability in place of × mark instead. And the value is the loss probability at the corresponding link when the node on the leftmost column functions as an ACD. When there are more than a path, the loss probability is calculated by utilizing the property in the 2nd assumption above.

Table 2 shows an example of the alarm loss probability chart based on the network of Figure 1. In this example, we have exaggerated the alarm loss probabilities and assumed that the probability increases by 0.1 in every hop distance from an ACD. The actual probabilities will be much smaller than those.

Table 2: Alarm loss probability chart.

	1	2	3	4	5	6	7	8
a	0.1^2	0.1	0.2	0.2				
b	0.1^2	0.2	0.1	0.1				
c	0.2		0.1			0.1		0.2
d	0.2	0.1		0.3				
e	0.2	0.3		0.1		0.1		0.1
f						0.2		0.1

Now we want to allocate the minimum number of ACDs keeping all the alarm loss probabilities over the network below a certain predefined level, called a *marginal alarm loss probability.*

We propose an algorithm for selecting the location of the ACDs:

1. Remove the probabilities greater than the marginal probability. Replace the remaining loss probabilities with × marks.
2. Below the rows obtained in the previous step, list rows composed of the combined rows whose elements become less than the marginal probability by combining those two rows, based on Assumption 3 above. This is performed repeatedly with regard to each column. If a combined row includes a row having an × mark in the current column the combined row should not be listed because the combined row is implied

by the × marked row found in the earlier step. When marking with × for the combined rows, the combined loss probabilities should be calculated from the the alarm loss probability chart. If × are copied directly from the upper rows, there is a chance for ×'s to be missed.

3. Continue Step 2 by combining three rows. The same cautions need to be applied as in Step 2. Repeat with more rows if necessary.

Example. Let the marginal alarm loss probability be 0.15 in the given example. First, single nodes which do not exceed the marginal probability are listed for each column. And then the combined nodes which do not exceed the marginal probability and do not include already listed nodes are listed for each column. For example, in column 1 even though the combined nodes *cd* are less than the marginal probability it is not included since *c* is already listed in the previous step, which means the alarm loss probability of *c* alone is already below the marginal probability. After following the steps above for the all columns, we have rows *a*, *b*, *cd*, *cd*, *de* for column 1; *a*, *d*, *be* for column 2; *b*, *c* for column 3; *b*, *e*, *ad* for column 4; *c*, *e* for column 5; *e*, *f* for column 6. We have three combined rows, i.e. *ad*, *be*, *de*.

Table 3 shows a link coverage chart considering the marginal loss probability. We note that the ×'s at $(ad, 4)$, $(be, 2)$, and $(de, 1)$ would be omitted if we simply copied them from the upper rows. By applying the same method used in simplifying the boolean function, we have a reduced chart as in Table 4. When reducing the chart, a caution must be taken so as not to delete a row by a row below having more nodes because the row having more literals, i.e., at the lower rows, increases the total number of nodes.

From Table 4 we find that two nodes, *b* and *e* are enough to cover all the critical links in the given network with the alarm loss probabilities less than or equal to 0.15.

Table 3: Link coverage chart given marginal prob.

	1	2	3	4	6	8
a	×	×				
b	×		×	×		
c	×		×		×	
d		×				
e				×	×	×
f						×
ad	×	×		×		
be	×	×	×	×	×	×
cd	×	×	×		×	
ce	×	×		×	×	×
de	×	×		×	×	×

Table 4: Reduced chart.

	2	3	8
a	×		
b		×	
c		×	
e			×
be	×	×	×

4 Provisioning Paths

After ACDs are deployed, paths may need to be changed depending on a demand pattern. In this section we investigate how to provision paths properly so that the alarms on the new path can be reported without being lost and the new path still can use the shortest one.

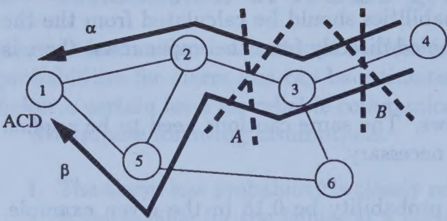


Figure 2: Multiple faults.

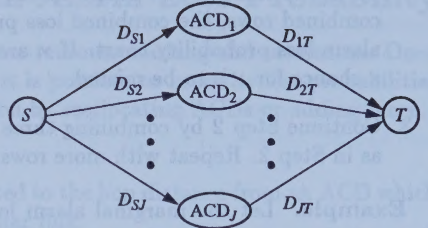


Figure 3: OD pair via ACDs.

Consider a multiple fault case as shown in Figure 2. When there are multiple faults at A and B, the alarms from B cannot be delivered due to another link failure at A. If path β was provisioned through 1 - 5 - 6 - 3 - 4, the alarm could reach the ACD. This example shows that if a new path does not share more than one link with paths built previously in the same direction, the probability for alarm loss can be reduced. This problem can be solved by letting an ACD be placed in the intermediate node of any shared links while the shortest path requirement is satisfied. We can achieve it in the following way: Suppose we want to provision a new shortest path from node S to T, so that any alarms due to multiple faults on the new path can be delivered to ACDs. Then we find the shortest paths from the origin to all ACDs, such as $D_{S1}, \cdots, D_{SJ}$ and the shortest paths from all ACDs to a destination, such as $D_{1T}, \cdots, D_{JT}$ under the constraint that a path should not share a link in the same direction more than once with other paths built earlier as shown in Figure 3. We choose the shortest path from the calculated path set, i.e. $\min_j[D_{Sj} + D_{jT}]$. Then, all the multiple failures will be reported to the multiple ACDs which are responsible for the failures without being blocked by one another.

Now, let us consider one of the two shortest paths, i.e., from node S to node ACD$_1$. We give priority to the paths deployed earlier so that they can not be changed. We assume that there always exists such a shortest path under the constraint stated above. We propose an algorithm to find those paths.

A shortest walk from a given node i to one of the ACDs, say ACD$_1$, subject to the following constraints is referred to as a *shortest* ($\leq h$) *walk* under the constraint $\mathcal{R}$ and its length is denoted by D_i^h:

1. The walk contains at most h arcs and goes through ACD$_1$ only once.

2. The walk should not share more than a link with another path built earlier in the same direction. (Let us denote this constraint as $\mathcal{R}$.)

Note that such a walk may not be a path, that is, it may contain repeated nodes; conditions under which this is not possible will be given later.

Let Q be the set of virtual path identifiers (VPIs) of all the paths currently provisioned in the network. Let J_i be the set of nodes connected directly to node i. We will prove shortly that D_S, the shortest walk from node S to ACD$_1$ satisfying the constraint $\mathcal{R}$, can be calculated by the equation

$$D_S = \min_i \left[d_{Si} + D_i\{p(S,i)\}\right], \tag{4}$$

where $p(S,i)$ is the VPI of the path deployed already on the arc (S,i), if any, and $D_i\{q\}$ is the distance associated with a shortest walk from i to 1 which does not share any link with VPI q *still* subject to the constraint $\mathcal{R}$. $D_i\{q\}$ can be found from the iteration

$$D_i^{h+1}\{q\} = \min_{j:\{q\}} \left[d_{ij} + D_j^h\{p(i,j)\}\right], \text{ for all } i \neq 1, \tag{5}$$

which means the minimum of $d_{ij} + D_j^h\{p(i,j)\}$ with varying i and without VPI q on its walk associated with $d_{ij} + D_j^h\{p(i,j)\}$.

By convention, we take

$$D_1^h\{q\} = 0, \text{ for all } h \text{ and all } q \in Q \tag{6}$$

and the algorithm starts from the initial condition

$$D_i^0\{q\} = \infty, \text{ for all } i \neq 1 \text{ and all } q \in Q. \tag{7}$$

The algorithm terminates after h iteration if $D_i^h\{q\} = D_i^{h-1}\{q\}$, for all i.

Proposition. Consider the algorithm (4) and (5) with the initial condition $D_i^0\{q\} = \infty$, for all $i \neq 1$. Then:

(a) The scalars D_i generated by the algorithm are equal to the shortest ($\leq h$) walk length from node i to node 1 under the constraint $\mathcal{R}$ that there be no more than one link shared by two walks in the same direction on the walk.

(b) The algorithm terminates after a finite number of iterations if and only if all cycles not containing ACD_1 have nonnegative length. Furthermore, if the algorithm terminates, it does so after at most $h \leq N - 1$ iterations, and at termination, D_i^h is the shortest path length from i to 1.

Proof: (a) We will only need to show $D_i^h\{q\}$ generated by (5) is equal to the shortest ($\leq h$) walk length from i to 1 without the path having VPI q on its walk while satisfying the constraint $\mathcal{R}$. It is because once we have the shortest ($\leq h$) walk, $D_i^h\{q\}$, for all the neighboring nodes of the originating end node S, then the shortest walk from node S to ACD_1 subject to the constraint $\mathcal{R}$ will be found trivially from (4).

We prove it by induction. From Equations (5), (6) and (7) we have for all $i \neq 1, S$,

$$D_i^1\{q\} = \min_{j:\{q\}} \left[d_{ij} + D_j^0\{p(i,j)\}\right] = \begin{cases} \infty, & \text{if } p(i,1) = q \\ d_{i1}, & \text{otherwise,} \end{cases}$$

since for any $q \in Q$

$$D_j^0\{q\} = \begin{cases} 0, & \text{if } j = 1 \\ \infty, & \text{otherwise,} \end{cases}$$

so $D_i^1\{q\}$ is indeed equal to the shortest (≤ 1) walk length from i to 1 without the provisioned path having VPI q on its walk under the constraint $\mathcal{R}$. Suppose that $D_i^h\{q\}$ is equal to the shortest ($\leq h$) walk length from i to 1 without the path having VPI q on its walk under the constraint $\mathcal{R}$. We will show that $D_i^{h+1}\{q\}$ is the shortest ($\leq h+1$) walk from i to 1 without VPI q on its walk while satisfying the constraint $\mathcal{R}$. Indeed, a shortest ($\leq h+1$) walk without VPI q from i to 1 either consists of less than $h+1$ arcs, in which case its length is equal to $D_i^h\{q\}$, or else it consists of $h+1$ arcs with the first arc being (i,j) for some $j \neq 1$, where $p(i,j) \neq q$, followed by an h-arc walk from j to 1.

The latter walk must be a shortest ($\leq h$) walk from j to 1 which does not have the same VPI on its walk as the one on arc (i,j), if any on (i,j). We thus have

$$\begin{aligned}&\text{Shortest } (\leq h+1) \text{ walk length without } q \text{ under } \mathcal{R}\\ &= \min\left\{D_i^h\{q\},\ \min_{j\neq 1:\{q\}}\left[d_{ij}+D_j^h\{p(i,j)\}\right]\right\},\end{aligned} \tag{8}$$

where $p(i,j) \neq q$. Using the induction hypothesis that $D_j^h\{q\}$, for $j \in J_i$, is the shortest ($\leq h$) walk without VPI $p(i,j)$ on its walk from j to 1, we have $D_j^h\{q\} \leq D_j^{h-1}\{q\}$, since the set of ($\leq h$) walks from node i to 1 contains the corresponding set of ($\leq h-1$) walks. Therefore,

$$D_i^{h+1}\{q\} = \min_{j:\{q\}}\left[d_{ij}+D_j^h\{q\}\right] \leq \min_{j:\{q\}}\left[d_{ij}+D_j^{h-1}\{q\}\right] = D_i^h\{q\}.$$

Furthermore, we have

$$D_i^h\{q\} \leq D_i^1\{q\} = \begin{cases}\infty, & \text{if } p(i,1)=q\\ d_{i1}, & \text{otherwise}\end{cases}$$

so from Equation (8) we obtain

$$\begin{aligned}&\text{Shortest } (\leq h+1) \text{ walk length without } q \text{ under } \mathcal{R}\\ &= \min\left\{D_i^h\{q\},\ \min_{j\neq 1:\{q\}}\left[d_{ij}+D_j^h\{p(i,j)\}\right]\right\}\\ &= \min\left\{D_i^h\{q\}, \min_{j:\{q\}}\left[d_{ij}+D_j^h\{p(i,j)\}\right]\right\} = D_i^{h+1}\{q\},\end{aligned}$$

where $D_1^h\{p(i,1)\} = 0$ is used. It completes the proof for the iteration (5).

(b) If the algorithm terminates after h iterations, we must have

$$D_i^k\{q\} = D_i^h\{q\}, \qquad \text{for all } i \text{ and } k \geq h \tag{9}$$

so we cannot reduce the lengths of the shortest walks by allowing more and more arcs in these walks. It follows that there cannot exist a negative-length cycle not containing ACD_1, since such a cycle could be repeated an arbitrarily large number of times in walks from some nodes to ACD_1, thereby making their length arbitrarily small and contradicting (9). Conversely, suppose that all cycles not containing ACD_1 have nonnegative length, then by deleting all such cycles from shortest ($\leq h$) walks, we obtain paths of less or equal length. Therefore, for every i and h, there exists a path that is a shortest ($\leq h$) walk from i to 1 without VPI q on its walk, and the corresponding shortest path length is equal to $D_i^h\{q\}$. Since paths have no cycles, they can contain at most $N-2$ arcs, excluding ACD_1 and node S. It follows that $D_i^{N-1}\{q\} = D_i^{N-2}\{q\}$, for all i implying that the algorithm terminates after at most $N-1$ iterations. **Q.E.D.**

A node i receives an array of distance values $[D_j^h\{\emptyset\}, D_j^h\{q_1\}, D_j^h\{q_2\}, \cdots, D_j^h\{q_{|Q|}\}]$ along with their corresponding shortest paths from one of its neighboring node, j, where $|Q|$ is the number of paths provisioned currently in the network. In fact each node does not have to know about what are in Q because each node can regard the VPIs which are not reported to it as non-existing VPIs.

Node i calculates the shortest distance D_i^h from Node i to ACD_1 using Equation (4) and the shortest distances without VPI q_1, VPI q_2, $\cdots$ and VPI $q_{|Q|}$ on its walk to ACD_1, i.e., $[D_i^h\{\emptyset\}, D_i^h\{q_1\}, D_i^h\{q_2\}, \cdots, D_i^h\{q_{|Q|}\}]$, by using Equation (5). If a shortest path is not available, we consider the distance as ∞.

Figure 4 shows the algorithm applied to a portion of the network shown in (a), where the ACD is at ACD_1 and the origin is at node 5. The VPI within braces indicates the path does not have that VPI on it. After all the pairs of the shortest paths from the origin to all ACDs and from all the ACDs to the destination are calculated, the path with minimum total length is selected.

5 Summary

We have investigated how to construct a reliable alarm structure in the communication networks. We have studied how to allocate a minimal set of ACDs and still cover all critical alarms in mesh networks. In order to find the minimal set we have represented the network topology and provisioned paths with matrices and used the link coverage chart. The probabilistic characteristic of alarm loss has been considered. For that case a new algorithm for a minimal set of ACDs has been proposed. We also have considered multiple faults for high level of reliable communication. We have investigated how to provision in mesh network paths while reducing alarm loss under the multiple fault condition in mesh networks.

References

[1] L. Ball. *Cost-Efficient Network Management.* McGraw-Hill, 1992.

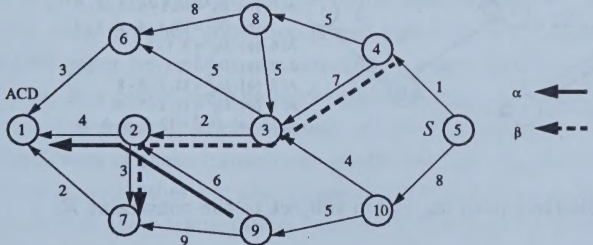
[2] S. Bibas, M.-O. Cordier, P. Dague, C. Dousson, F. Lévy, and L. Rozé. Alarm driven supervision for telecommunication network: I – off-line scenarios generation. *Annales des Telecommunications*, 51(9–10):493–500, 1996.

[3] Z. Kohavi. *Switching and Finite Automata Theory.* McGraw-Hill, 2nd edition, 1970.

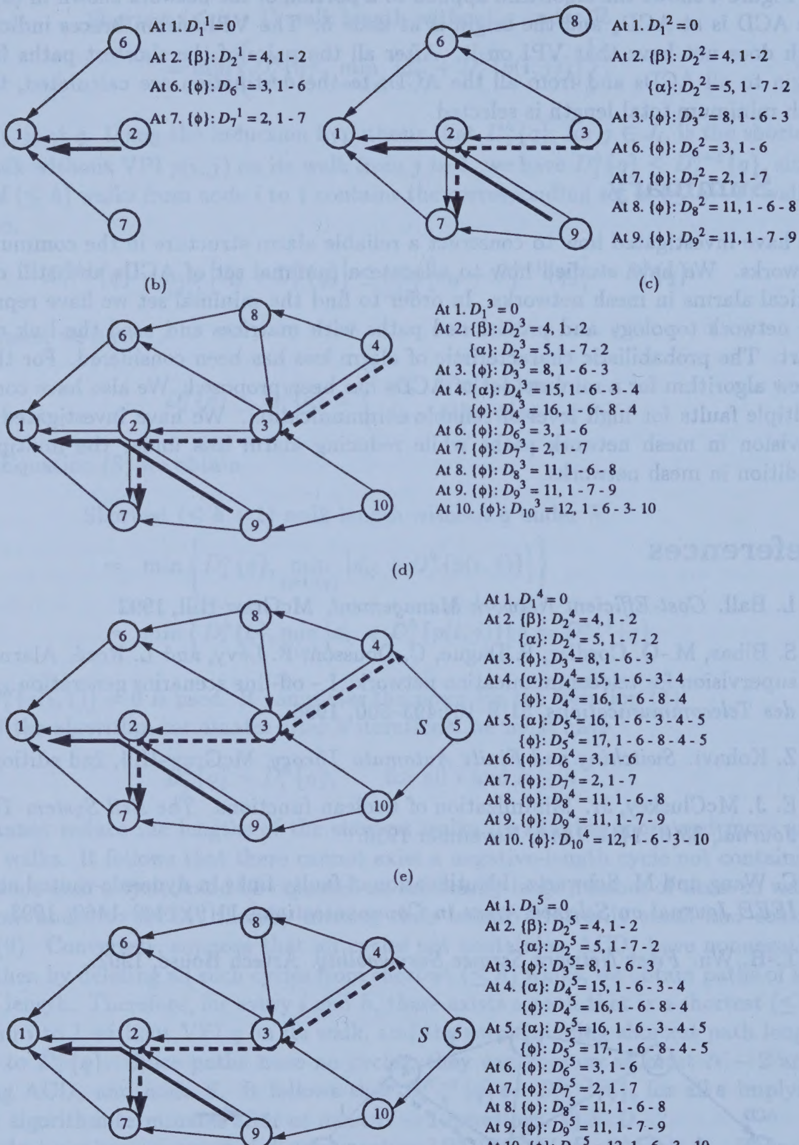
[4] E. J. McCluskey, Jr. Minimization of boolean functions. *The Bell System Technical Journal*, 35(6):1417–1444, November 1956.

[5] C. Wang and M. Schwartz. Identification of faulty links in dynamic-routed networks. *IEEE Journal on Selected Areas in Communications*, 11(9):1449–1460, 1993.

[6] T.-H. Wu. *Fiber Network Service Survivability.* Artech House, 1992.



(a)



At 1. $D_1^1 = 0$
At 2. $\{\beta\}$: $D_2^1 = 4$, 1 - 2
At 6. $\{\phi\}$: $D_6^1 = 3$, 1 - 6
At 7. $\{\phi\}$: $D_7^1 = 2$, 1 - 7

(b)

At 1. $D_1^2 = 0$
At 2. $\{\beta\}$: $D_2^2 = 4$, 1 - 2
$\{\alpha\}$: $D_2^2 = 5$, 1 - 7 - 2
At 3. $\{\phi\}$: $D_3^2 = 8$, 1 - 6 - 3
At 6. $\{\phi\}$: $D_6^2 = 3$, 1 - 6
At 7. $\{\phi\}$: $D_7^2 = 2$, 1 - 7
At 8. $\{\phi\}$: $D_8^2 = 11$, 1 - 6 - 8
At 9. $\{\phi\}$: $D_9^2 = 11$, 1 - 7 - 9

(c)

At 1. $D_1^3 = 0$
At 2. $\{\beta\}$: $D_2^3 = 4$, 1 - 2
$\{\alpha\}$: $D_2^3 = 5$, 1 - 7 - 2
At 3. $\{\phi\}$: $D_3^3 = 8$, 1 - 6 - 3
At 4. $\{\alpha\}$: $D_4^3 = 15$, 1 - 6 - 3 - 4
$\{\phi\}$: $D_4^3 = 16$, 1 - 6 - 8 - 4
At 6. $\{\phi\}$: $D_6^3 = 3$, 1 - 6
At 7. $\{\phi\}$: $D_7^3 = 2$, 1 - 7
At 8. $\{\phi\}$: $D_8^3 = 11$, 1 - 6 - 8
At 9. $\{\phi\}$: $D_9^3 = 11$, 1 - 7 - 9
At 10. $\{\phi\}$: $D_{10}^3 = 12$, 1 - 6 - 3 - 10

(d)

At 1. $D_1^4 = 0$
At 2. $\{\beta\}$: $D_2^4 = 4$, 1 - 2
$\{\alpha\}$: $D_2^4 = 5$, 1 - 7 - 2
At 3. $\{\phi\}$: $D_3^4 = 8$, 1 - 6 - 3
At 4. $\{\alpha\}$: $D_4^4 = 15$, 1 - 6 - 3 - 4
$\{\phi\}$: $D_4^4 = 16$, 1 - 6 - 8 - 4
At 5. $\{\alpha\}$: $D_5^4 = 16$, 1 - 6 - 3 - 4 - 5
$\{\phi\}$: $D_5^4 = 17$, 1 - 6 - 8 - 4 - 5
At 6. $\{\phi\}$: $D_6^4 = 3$, 1 - 6
At 7. $\{\phi\}$: $D_7^4 = 2$, 1 - 7
At 8. $\{\phi\}$: $D_8^4 = 11$, 1 - 6 - 8
At 9. $\{\phi\}$: $D_9^4 = 11$, 1 - 7 - 9
At 10. $\{\phi\}$: $D_{10}^4 = 12$, 1 - 6 - 3 - 10

(e)

At 1. $D_1^5 = 0$
At 2. $\{\beta\}$: $D_2^5 = 4$, 1 - 2
$\{\alpha\}$: $D_2^5 = 5$, 1 - 7 - 2
At 3. $\{\phi\}$: $D_3^5 = 8$, 1 - 6 - 3
At 4. $\{\alpha\}$: $D_4^5 = 15$, 1 - 6 - 3 - 4
$\{\phi\}$: $D_4^5 = 16$, 1 - 6 - 8 - 4
At 5. $\{\alpha\}$: $D_5^5 = 16$, 1 - 6 - 3 - 4 - 5
$\{\phi\}$: $D_5^5 = 17$, 1 - 6 - 8 - 4 - 5
At 6. $\{\phi\}$: $D_6^5 = 3$, 1 - 6
At 7. $\{\phi\}$: $D_7^5 = 2$, 1 - 7
At 8. $\{\phi\}$: $D_8^5 = 11$, 1 - 6 - 8
At 9. $\{\phi\}$: $D_9^5 = 11$, 1 - 7 - 9
At 10. $\{\phi\}$: $D_{10}^5 = 12$, 1 - 6 - 3 - 10

(f)

Figure 4: Shortest path algorithm subject to the constraint $\mathcal{R}$.

On Approximation Algorithms for Radio Network Scheduling

Arunabha Sen
Department of Computer Science and Engg.
Arizona State University
Tempe, AZ 85287
USA

Ewa Malesinska
Department of Mathematics
Technical University of Berlin
Berlin D-10623
GERMANY

Abstract: The optimal broadcast schedule construction problem for a radio network is known to be NP-Complete. In this paper we present an approximation algorithm for the problem with a guaranteed performance bound.

1 Introduction

A packet radio network is a collection of radio transmitters and receivers (transceivers) located in a geographical region. Associated with each transceiver is a transmission range, which depends on its transmission power. A transceiver T_j located within the transmission range of another transceiver T_i can receive the messages from T_i. The transceivers of the network share the communication medium for message transfer. This gives rise to conflict situations where two or more of the transceivers may want to use the same shared medium simultaneously. Such conflicts result in garbling and eventual loss of messages. Communication protocols are used to avoid such events. Such protocols may be divided into two broad classes, *static* and *dynamic.* In case of a *static* protocol an a priori assignment of the communication resources, such as time, frequency or code, is made among the transceivers to avoid contention. Dynamic protocols have no a priori resource allocation, instead resource requests are met on demand. *Time Division*, *Frequency Division* and *Code Division* multiple access (referred to as TDMA, FDMA and CDMA respectively) are examples of the a priori assignment of resources. ALOHA and CSMA are examples of dynamic protocols. Static allocation is well suited for a steady stream of traffic while dynamic allocation is well suited for bursty traffic .

In this paper we restrict our attention to static allocation in general and TDMA in particular. In a packet radio network operating under a TDMA scheme, some transceivers are allowed to transmit simultaneously. However, the geographic locations and the transmitting powers of these transceivers should be such that not only do they not cause interference to each other but also do not cause interference at any other transceiver which may be within their transmission ranges. A *schedule* is a sequence of fixed length time slots where a slot is assigned to a set of non-interfering transceivers for transmission of messages. *Schedule length* is measured by the number of time slots in a schedule. A schedule is said to be *optimal* if it uses the minimum number of slots.

In [5], [6] it was shown that the optimal schedule construction problem is NP-complete. In this paper we present an approximation algorithm for the problem with a constant performance bound. We show that the optimal schedule construction problem remains NP-complete, even when the transmission range of all the transceivers are equal. We also present two approximation algorithms for this special case (one offline and one online) with constant performance bound.

2 Complexity of Broadcast Scheduling

In the radio network scheduling problem there could be two types of interferences: (i) if two transceivers a and b are within the transmission range of one another and they transmit at the same time, then they will cause interference to each other, (ii) if the transceivers a and b are not within the transmission range of one another, but there is a third transceiver c which is within the transmission range of both a and b, then simultaneous transmissions from a and b will cause interference at c. These two types of interferences are referred to as the type 1 and the type 2 respectively.

For the purpose of constructing an optimal schedule all prior research modeled the radio network as a graph, where a node represents a transceiver and there is a directed edge from node v_i to v_j if the transceiver j is within the transmission range of transceiver i. It may be noted that the resulting graph is a directed graph in which the existence of an edge from v_i to v_j does not necessarily imply the existence of an edge from v_j to v_i.

Most of the earlier research, [2], assumed that the graph constructed following the above rules will be an arbitrary graph. The optimal schedule construction problem was formulated as the problem of coloring the nodes of this graph using the fewest number of colors such that two nodes have the same color if they have a path of length 1 or 2. This coloring problem was shown to be NP-complete in [2]. However, Sen and Huson in [5], [6] showed that the graph constructed following the above construction rules will not be any arbitrary graph but a special class of graphs. Since the NP-completeness result of a problem in a certain domain does not necessarily mean that the problem will remain NP-complete when the domain is substantially restricted, the NP-completeness results of the network scheduling problem with an arbitrary graph as the model [2] does not accurately reflect the complexity of the problem.

The notion of *Planar point graphs* was introduced in [5], [6] for the purpose of modelling the radio networks. Given a set of points $P = (p_1, ..., p_n)$ in an m-dimensional space, each specified by an m vector $(x_1, ..., x_m)$ and a range r_i associated with each p_i, the *point graph* constructed from this set of points is a graph $G = (V, E)$, where each node v_i represents a point p_i and there is a directed edge from node v_i to v_j if $d(p_i, p_j) \leq r_i$, where $d(p_i, p_j)$ is the Euclidean distance between the points p_i and p_j. *Planar point graphs* are point graphs constructed from a set of points on a two dimensional plane ($m = 2$).

The optimal broadcast schedule construction (OBSC) problem can be stated in terms of coloring the nodes of a planar point graph. As discussed earlier, any broadcast schedule must be free of interference of type 1 and 2. In terms of the model, the problem can be stated as follows:

Optimal Broadcast Schedule Construction (OBSC) Problem

INSTANCE: Given a planar point graph $G=(V, E)$ and an integer constant k.

QUESTION: Is there a partition $V=V_1 \cup \ldots \cup V_k$ such that if u and v are distinct nodes of some set V_i in the partition then (i) there is no edge from u to v, (ii) there is no edge from v to u and (iii) there is no node w in the graph G such that there is an edge from u to w and from v to w?

We will also refer to the OBSC problem as either *distance-2 k-colorability of planar point graphs* or *distance-2 coloring of planar point graphs.*

In [5], [6] it was shown that the OBSC problem is NP-complete by proving the following theorem.

Theorem 1 *The distance-2 3-colorability of planar point graphs is NP-complete [5], [6].*

3 Approximation Algorithm for Distance-2 Coloring of Planar Point Graphs

In this section we present an approximation algorithm for the distance-2 coloring of planar point graphs with a constant performance bound.

The following theorem is due to Szekeres and Wilf in [7].

Theorem 2 *Any undirected graph* $G = (V, E)$ *can be colored using at most* $\delta(G) + 1$ *colors (so that no two adjacent nodes have the same color), where* $\delta(G)$ *denotes the largest* ρ *such that* G *contains a subgraph in which every vertex has a degree at least* ρ.

Matula and Beck in [4] proposed the *smallest-last coloring* algorithm for coloring the nodes of an undirected graph $G = (V, E)$, which can be implemented in $O(|V| + |E|)$ time. It was also shown in [4] that this algorithm will not need more than $\delta(G)+1$ colors.

Our approximation algorithm for distance-2 coloring of a planar point graph first constructs an undirected graph from the planar point graph (which is directed). This graph will be referred to as the *Derived Graph.* Given a planar point graph $G = (V, E)$, the corresponding derived graph $H = (V, E')$ is constructed as follows: The node set for the graph H is the same as the node set of graph G. There is an edge between the nodes u and v in H, if in G at least one of the following three conditions is satisfied: (i) there is directed edge from u to v, (ii) there is a directed edge from v to u, (iii) there is a node w, such that there is a directed edge from u to w and v to w in G.

From the construction of the derived graph, it is clear that the distance-2 chromatic number of the planar point graph G is the same as the distance-1 chromatic number (or just the chromatic number) of the derived graph H. Therefore instead of trying to compute the distance-2 chromatic number of the planar point graph, we try to compute the chromatic number of H. We apply the smallest-last algorithm to color the nodes of H. From theorem 2, it is known that the number of colors used by this algorithm is bounded by $\delta(H)+1$. Let $APPCOLOR(H)$ denote the number of colors needed to color the nodes of H using this algorithm. Then

$$APPCOLOR(H) = \delta(H) + 1. \tag{1}$$

Let $OPTCOLOR(H)$ represent the number of colors needed to optimally color the nodes of H. We claim the following relation between the $OPTCOLOR(H)$ and $\delta(H)$:

Claim 1: $OPTCOLOR(H) \geq \delta(H)/14 + 1$

Proof: Let H' be a subgraph of H such that every vertex in H' has a degree at least $\delta(H)$. The nodes of H' represent transceivers in specific locations in the plane. If u and v are nodes of H (and hence of G), we use the notation $T(u)$ to indicate the transceiver associated with node u, $R(u)$ to indicate the transmission range of $T(u)$, and $D(u,v)$ to represent the Euclidean distance between $T(u)$ and $T(v)$. Suppose that the node v^* of H' corresponds to the transceiver with the smallest transmission radius. Let the set of adjacent nodes of v^* in H' be denoted by $N_{H'}(v^*)$. By our choice of H',

$$|N_{H'}(v^*)| \geq \delta(H). \tag{2}$$

The set of adjacent nodes of v^* in H' is divided into three types:

Type-A: A node $u \in N_{H'}(v^*)$ is a Type-A neighbor if $R(v^*) \geq D(v^*, u)$.

Type-B: A node $u \in N_{H'}(v^*)$ is a Type-B neighbor if $R(u) \geq D(v^*, u)$.

Type-C: A node $u \in N_{H'}(v^*)$ is a Type-C neighbor if there is node w in H, such that $R(v^*) \geq D(v^*, w)$ and $R(u) \geq D(u, w)$. It may be noted that w may or may not be a node of H'.

The types are prioritized in the sense that if a node satisfies both type A and B conditions, it will be classified as a type A neighbor.

To prove our claim, we try to find the minimum number of sectors in which the region surrounding $T(v^*)$ can be divided (as shown in figure 1), so that the nodes of H' that represent the transceivers in a sector form a clique in H'.

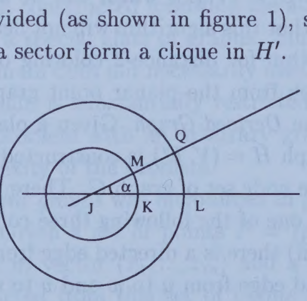


Figure 1:

We assume that $R(v^*)$ is r_{v^*}. We draw $T(v^*)$ and the associated transmission circle of radius r_{v^*} as shown in figure 1. We draw another circle of radius $2r_{v^*}$ with center at the location of $T(v^*)$. The location of $T(v^*)$, i.e., the center of the circles is represented as point J in the figure. Draw a line through J towards the circumference of the circles and suppose that it meets the smaller circle at K and the larger circle at L. Draw a second line through J towards the circumference of the circles and suppose that this line meets the smaller circle at M and the larger circle at Q. Supppose the angle between the lines JKL and JMQ is α. We would like to find the largest value of α such that the subset of nodes of $N_{H'}(v^*)$, representing the transceivers in the sector bounded by the lines JKL and JMQ, form a clique in H'.

Suppose the points L and Q are located in such a way that the distance between them is $r_{v^*} + \epsilon$, where ϵ is a very small number in comparison to r_{v^*}. Now, consider two transceivers at points L and Q with transmission radii r_{v^*} each. Also two transceivers at points K and M with transmission radii 0. Now in the graph H', both the nodes representing the transmitters at L and Q will be adjacent to node v^*, but they will not be adjacent to each other. However, if the distance between L and Q is reduced to r_{v^*}, then the corresponding nodes must be adjacent to each other.

If the distance between L and Q is r_{v^*}, then the angle α is $2sin^{-1}1/4$ which is about 28^o. Next we show that the nodes which represent transceivers in this sector and also adjacent to the node v^* in H', form a clique in H'.

Consider the subset of $N_{H'}(v^*)$ comprised of only the Type A neighbors. Clearly, transmission from the transmitters corresponding to these nodes reaches $T(v^*)$. Therefore every pair of nodes in this subset will be mutually adjacent in H and thereby form a clique. Now among the remaning nodes in $N_{H'}(v^*)$, consider the subset of $N_{H'}(v^*)$ that represents the transceivers in the sector bounded by lines JKL and JMQ with $\alpha = 28^o$. This subset of nodes can be divided into Type B and C neighbors as defined earlier. Consider two nodes u and v representing two transceivers in this sector. We show that u and v are adjacent by considering three different cases:

Case I: $(u, v \in \ Type-B)$ Suppose that $D(v, v^*) \geq D(u, v^*)$. Since transmission from $T(v)$ reaches $T(v^*)$, it must also reach $T(u)$ which is located in the 28^o sector bounded by the lines JKL and JMQ. As such u and v must be adjacent in H'.

Case II : $(u \in \ Type - B, v \in \ Type - C)$ Since v is a Type-C neighbor of v^*, there must be a node w in H such that $T(w)$ is within the transmission radius of both $T(v^*)$ and $T(v)$. Since $T(u)$ is a Type-B neighbor of v^*, transmission from $T(u)$ reaches $T(v^*)$ and as such it must also reach $T(w)$. This implies u and v must be adjacent in H'.

Case III: $(u, v \in \ Type - C)$ Since both u and v are Type-C neighbors of v^*, there must be nodes w_1 and w_2 in H such that $T(w_1)$ is within the tranmission radius of both $T(v^*)$ and $T(u)$, and $T(w_2)$ is within the transmission radius of both $T(v^*)$ and $T(v)$. From our choice of v^*, it is also known that the transmission radii of $T(u)$ and $T(v)$ is at least r_{v^*}. Suppose that between $T(u)$ and $T(v)$, $T(u)$ has the smaller transmission radius. Since a larger transmission radius of $T(u)$ will cover a larger area of the sector, we assume that $T(u)$ has the smallest possible radius, which is r_{v^*}. First we try to find a possible location for $T(v)$ such that the nodes u and v are not adjacent. If $T(u)$ is located in the position shown in figure 2, then there are two potential locations for $T(v)$ - region A and region B.

The region A will be empty if the position of $T(u)$ is moved closer to $T(v^*)$, as is

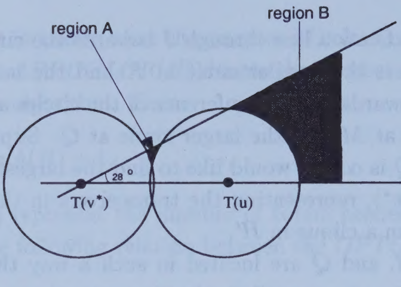


Figure 2:

shown in figure 3. From figure 3, $D(T(u), Q) = 2r_{v^*} cos\ 28^o - r_{v^*} = .77r_{v^*}$. Accordingly, we consider two different subcases: (i) $D(u, v^*) \leq 1.77r_{v^*}$ and (ii) $D(u, v^*) > 1.77r_{v^*}$.

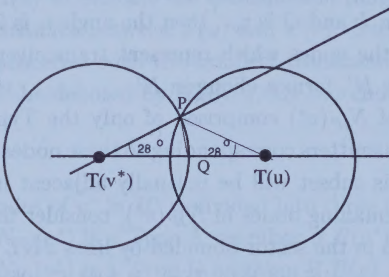


Figure 3:

The subcase (i) situation is shown in figure 4, where $T(u)$ is shown just outside the transmission radius of $T(v^*)$. The transmission area covered by $T(u)$ is shown as the circle with radius r_{v^*} and center at $T(u)$.

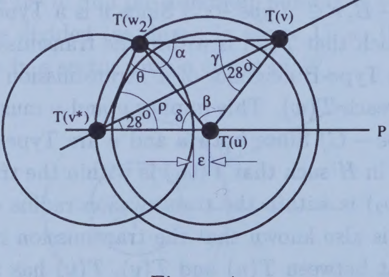


Figure 4:

If $T(w_2)$ is within this area, then u and v are adjacent. The only way u and v may not be adjacent, is if $T(w_2)$ is outside this circle. Suppose $T(u)$ is located ϵ distance away from the circumference of the transmission circle of $T(v^*)$ as shown in figure 4. Draw a line passing through $T(v^*)$ and $T(u)$ and extend it to some point P. This is one of the bounding lines of the 28^o sector. Draw a line with a 28^o slope with the line $T(v^*)T(u)P$.

Draw a circle of radius $r_{v^*} + \epsilon$ with $T(u)$ as the center. Suppose that the intesection point of the 28^o slope line with this circle is the location of $T(v)$ and intersection point of this circle and the transmission circle of $T(v^*)$ is the location of the transceiver $T(w_2)$, as shown in figure 4. Since $D(u, v)$ and $D(u, w_2)$ are both $r_{v^*} + \epsilon$, transmission from $T(u)$ does not reach v and w_2, as its transmission range is only r_{v^*}. $T(v)$'s transmission range should be large enough to reach $T(w_2)$. Next we will show that $D(v, w_2) > D(v, u)$ and as such transmission from $T(v)$ will reach $T(u)$ making the nodes u and v adjacent in H'.

The angles $\alpha, \ldots, \rho$ are as shown in figure 4. As in the triangle formed by $T(v^*)T(u)T(w_2)$, $D(u, v^*) = D(u, w_2)$, $\rho = \theta$. Similarly in the triangle formed by $T(v)T(u)T(w_2)$, $D(u, v) = D(u, w_2)$, $\alpha = \gamma$. As in the triangle formed by $T(v^*)T(u)T(w_2)$, $D(v^*, u) = r_{v^*} + \epsilon > r_{v^*} = D(v^*, w_2)$, $\delta < \theta$. This together with the fact that $\rho = \theta$ leads us to conclude that $\delta = 60^o - \eta$, where η is a small number.

As in the triangle $T(v^*)T(u)T(v)$, $D(u, v^*) = D(u, v)$, the angle $T(v^*)T(v)T(u) = T(v)T(v^*)T(u) = 28^o$. This implies the angle $T(v)T(u)P$ is 56^o. Therefore, $\beta = 180^o - 56^o - 60^o + \eta = 64^o + \eta$. This leads us to conclude that $\alpha < \beta$, and $D(v, w_2) > D(v, u)$. Since transmission from $T(v)$ reaches $T(w_2)$, $R(v)$ must be large enough so that it must also reach $T(u)$, making u and v adjacent. It is easy to verify that if $T(v)$ is located at a point different from the one considered above, the difference between $D(v, w_2)$ and $D(v, u)$ will be even larger.

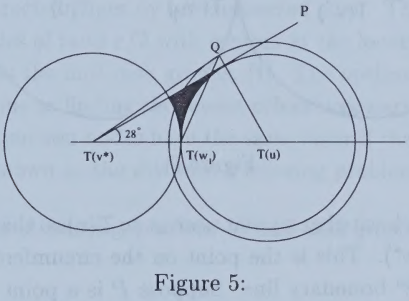


Figure 5:

The subcase (ii) situation is shown in figure 5, where $T(u)$ is shown as far outside the transmission radius of $T(v^*)$ as possible. Since the transmission radius of $T(u)$ is assumed to r_{v^*}, its location must be within a circle of radius $2r_{v^*}$ with the center at $T(v^*)$. The transmission area covered by $T(u)$ is shown as the circle with radius r_{v^*} and center at $T(u)$. Also the location of $T(w_1)$ in this subcase must be on the circumference of the transmission circle of $T(v^*)$, as shown in figure 5. We assume that $T(v)$ is outside the transmission circle of $T(u)$, otherwise u and v would be adjacent.

We consider two possible locations for $T(v)$. First we show that if $T(v)$ is located in the area enclosed by the circumference of two circles and the 28^o boundary line $T(v^*)P$ (shown as the shaded area in figure 5), its transmission will reach $T(w_1)$ and u and v will be adjacent. In order to show this, we first draw a circle of radius $r_{v^*} + \epsilon$ with $T(u)$ as the center. Find a point Q on the circumference of this circle such that $D(w_1, Q) = r_{v^*}$. Draw lines $T(v^*)Q$, $T(w_1)Q$ and $T(u), Q$. In the triangle $T(w_1)T(u)Q$, $D(w_1, Q) = D(w_1, u) = r_{v^*}$ and $D(u, Q) = r_{v^*} + \epsilon$. Therefore in the triangle $T(w_1)T(u)Q$, the angle $T(u)T(w_1)Q$ is $60^o + \eta$ where η is a small number. Therefore in the triangle $T(v^*)T(w_1)Q$, the angle $T(v^*)T(w_1)Q$ is $120^o - \eta$. Since $D(v^*, w_1) = D(w_1, Q) = r_{v^*}$, both the angles $T(w_1)T(v^*)Q$ and $T(w_1)QT(v^*)$ must be $30^o + \eta/2$. Since $T(v)$ is located within the 28^o boundary line, its distance to $T(w_1)$ must be less than the distance of the point Q to $T(w_1)$, which is r_{v^*}. Since the transmission range of $T(v)$ is at least r_{v^*}, its transmision must reach $T(w_1)$, making the nodes u and v adjacent in H'.

In the other possibility, suppose that $T(v)$ is located in the 28^o sector outside the range of $T(u)$, away from the transmission range of $T(v^*)$ as shown in figure 6. In this case we will show that $D(v, w_2) > D(v, u)$ so that transmission from v will reach u, making the nodes u and v adjacent in H'.

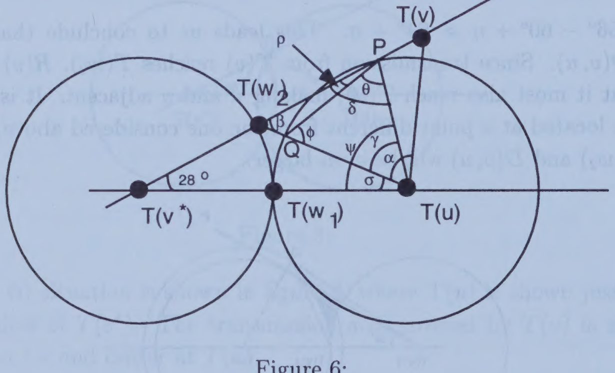


Figure 6:

Suppose that $T(w_2)$is located at a point nearest to $T(v)$ so that it receives transmission from both $T(v)$ and $T(v^*)$. This is the point on the circumference of the transmission circle of $T(v^*)$ on the 28^o boundary line. Suppose P is a point on the circumference of the transmission circle of $T(u)$ such that $D(v^*, P) = 2r_{v^*}$. Draw lines $T(u)P$, $T(u)T(v)$ and $T(u)T(w_2)$. Suppose the line $T(u)T(w_2)$ meets the circumference of the transmission circle of $T(u)$ at the point Q. Draw a line PQ.

The angles $\alpha, \beta, \ldots$ are shown in the figure 6. If $\alpha > \beta$ then $D(v, w_2) > D(v, u)$. $\psi = \delta = 1/2(180^o - 28^o) = 76^o$. Since $D(v^*, w_2) < D(u, w_2)$, $\sigma < 28^o$. Therfore $\gamma = \psi - \sigma > 48^o$, and consequently $\alpha > 48^o$. Since $\gamma > 48^o$ this implies $\phi = \theta < 1/2(180^o - 48^o) = 61^o$. Therefore, $\rho = \delta - \theta > 76^o - 61^o = 15^o$. Since $\phi < 61^o$ and $\rho > 15^o$, $\beta = \phi - \rho < 46^o$. Since $\alpha > 48^o$ and $\beta < 46^o$, $D(v, w_2) > D(v, u)$. Therefore transmission from $T(v)$ must reach $T(u)$, making the nodes u and v adjacent in H'.

Thus if $\alpha = 28^o$, in all cases, the subset of nodes of $N_{H'}(v^*)$, representing the transceivers in the sector defined by α, form a clique in H'.

The region around $T(v^*)$ can be divided into $\lceil 360^o/28^o \rceil = 13$ sectors. The subset of Type B and C neighbors of $N_{H'}(v^*)$ representing the transceivers in each of the sectors form a clique in H'. As discussed earlier, the Type A neighbors form a clique all by themselves. Therefore, the size of the largest independent set in $N_{H'}(v^*)$ is 13+1. Thus any valid coloring (no two adjacent nodes have the same color) of the subgraph induced by the set $N_{H'}(v^*) \cup \{v^*\}$ must use at least $N_{H'}(v^*)/14 + 1$ colors. This implies

$$OPTCOLOR(H) \geq N_{H'}(v^*)/14 + 1 \tag{3}$$

Combining (2) and (3) we prove claim 1.

$$OPTCOLOR(H) \geq \delta(H)/14 + 1 \tag{4}$$

Combining (1) and (4) we have the following theorem.

Theorem 3 *The ratio between the performance of the approximation algorithm for distance-2 coloring of a planar point graph to the optimal algorithm is bounded by 14.*

4 Complexity Results when the Transmission Radii are Equal

When the transmission range r of all the transceivers are equal, then in the graph we replace two oppositely directed edges by an undirected edge. This graph is same as the intersection graph of circles of radii $r/2$ with centers at the locations of the transceivers. Such graphs are known as the unit disk graphs, [1]. The optimal schedule construction problem in this case is same as finding the fewest colors necessary to color the nodes of a unit disk graph such that no two nodes have the same color if they have a path of length 1 or 2. This problem is known as the distance-2 coloring problem of unit disk graphs.

Theorem 4 *The distance-2 coloring problem of unit disk graphs is NP-complete, even for fixed $k = 3$.*

Proof: The proof uses similar ideas as in [1]. The theorem is proved by sketching a polynomial time transformation from the *planar graph 3-colorability with node degree at most 4* problem, which was shown to be NP-complete in [3]. From a planar graph G with node degree at most 4 we generate a unit disk graph G' and show that G is 3-colorable iff G' is distance-2 3-colorable. The details of the proof is left out due to the limitations on the number of pages.

5 Conclusion

In this paper we presented some approximation algorithms for the problems of radio network scheduling which are known to be NP-complete. In the abstract model the problems are equivalent to the distance-2 coloring problem of planar point graphs and unit disk graphs respectively. For both the cases, we presented approximation algorithms with a constant performance bound. This is contrast to the general graph coloring problem where it is known that no polynomial time algorithm can provide a performance bound of $O(n^\epsilon)$ for some $\epsilon > 0$ unless $\mathbf{P} = \mathbf{NP}$.

Acknowledgement: The authors wish to express their sincere thanks to Professor Errol Lloyd and Dr. Madhav Marathe for pointing out errors in an earlier version of the paper.

References

[1] B.N. Clark, C.J. Colbourn and D.S. Johnson, Unit disk graphs, *Discrete Mathematics* 86 (1990), pp. 165-177.

[2] A. Ephremides and T.V. Troung, Scheduling broadcasts in a radio network, *IEEE Trans. on Communications*, vol. 38 (1990), pp. 456-461.

[3] M.R. Garey, D.S. Johnson and L.J. Stockmeyer, Some simplified np-complete graph problems, *Theoretical Computer Science*, vol. 1 (1976), pp. 237-267.

[4] D.W. Matula and L.L. Beck, Smallest-last ordering and clustering and graph coloring algorithms, *Journal of the Association of Computing Machinery* 30 (1983), pp. 417-427.

[5] A. Sen and M. Huson, A new model for scheduling packet radio networks in *Proceedings of IEEE INFOCOM'96*, pp. 1116-1124.

[6] A. Sen and M. Huson, A new model for scheduling packet radio networks, *Wireless Networks*, vol. 3 (1997) no. 1, pp. 71-82.

[7] G. Szekeres and H.S. Wilf, An inequality for the chromatic number of a graph, *J. of Combinatorial Theory* 4 (1968), pp. 1-3.

Effect of Packet Lengths and Power Levels on Random Multiple Access *

Wei Luo
Electrical Engineering Department
University of Maryland
College Park, MD 20742

Anthony Ephremides
Electrical Engineering Department
University of Maryland
College Park, MD 20742

To enhance the throughput of slotted random access in a radio communication system, multiple power levels have been proposed [1-4]. For example, in [4], it is proposed that the transmitters in an ALOHA system randomly choose one of m discrete power levels during each transmission attempt. The use of the variation of the received power levels makes capture possible. In [4], the optimization of the values of the power levels is investigated given that their number is m and that they must be in the interval $[P_{min}, P_{max}]$.

We observe that for a packet which is sent at a higher power level, the duration of the symbols in the packet can be reduced since the required levels of E_s/N_0 can be met. By reducing the duration of the symbols and, hence, the packet lengths, we can reduce the chance of packets collision. On the other hand, if all the users use the maximum possible power, the likelihood of packet overlap(and hence, collision) decreases, but the possibility of capture(with its beneficial effects on throughput) is eliminated. In this paper, we investigate the tradeoff between reducing the packet length (by increasing the transmission power to meet the requirement of E_s/N_0) and determine the number and values of different power levels in order to retain the capture advantage. At the same time, we note that for mobile wireless applications, energy efficiency is of paramount importance. Therefore, since random access schemes involve collisions and retransmissions, it is possible to waste energy on repeated transmissions of the same packet. Thus we also consider the throughput under the constraint of system power consumption as the relevant cost function.

Our model is based on the following assumptions:

A1 All packets have the same number of symbols M.

A2 The receiver always pick the strongest packet to decode. Only if the packet is the strongest among those of the colliding packets and its SNIR is greater than a fixed threshold can it be successfully received.

A3 Power level can not exceed the maximum available power P_{max}.

A4 The arrival times of the offered packets(including newly arrival packets and retransmission packets) is assumed Poisson [3, 5] and the channel is the usual collision channel.

---

*Prepared through collaborative participation in the Advanced Telecommunications/Information Distribution Research Program (ATIRP) Consortium sponsored by the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0002.

A5 No multiuser detection is assumed and a packet is correctly received if all its symbols are correctly decoded.

If two packets are transmitted respectively at power levels P_i and P_j ($P_i \geq P_j$), then the packet at level P_i will be received correctly only if

$$\frac{P_i L_i}{P_{i+1} \min(L_i, L_{i+1}) + N_0} \geq MT_{E_s/N_0} \triangleq C \tag{1}$$

Let G_i be the offered traffic rate at the ith power level. The total throughput is given by

$$S(\mathbf{G}) \leq \sum_{i=1}^{m} G_i \exp[-\sum_{j=1}^{i} G_j \max(L_i, L_j)] \tag{2}$$

Equality holds in Inequality(2) when $m = 1$, i.e. when a single power level is used.

Based on (1) and (2), we prove analytically that *if* $C > e - 1 \approx 1.718$, *use of single power level always achieves the maximum possible throughput.* In such single power level systems, the common power level is equal to P_{max} and the packet length is as short as is necessary to create $E_s/N_0 \geq C$, i.e. $L = CN_0/P_{max}$.

If the overall average system power is constrained by $P_x = \sum_{i=1}^{m} G_i P_i L_i \leq P_s$,, the maximum throughput becomes a function of P_s, denoted by $S_{max}(P_s)$. We prove that *use of single power level achieves the largest possible throughput* $S_{max}(P_s)$ *for any value of* P_s, *if* $C > e - 1$. [1]

References

[1] I.Cidon J.Kodesh and M.Sidi, "Erasure, capture, and random power level selection in multipleaccess systems," *IEEE Trans. Commun.*, vol. 36, pp. 263–271, Mar. 1988.

[2] C.C. Lee, "Random signal levels for channel access in packet broadcast networks," *IEEE J. Select. Areas Commun.*, vol. 5, pp. 1026–1034, July 1987.

[3] J.J. Metzner, "On improving utilization in ALOHA networks," *IEEE Trans. Commun.*, pp. 447–448, April 1976.

[4] R.O.LaMaire A.Krishna and M.Zorzi, "Optimization of capture in mulitple access radio systems with rayleigh fading and random power levels," in *Multiaccess, Mobility and Teletraffic for Personal Communications*, B.Jabbari P.Godlewski and X.Lagrange, Eds., pp. 321–336. Boston,MA:Kluwer Acadenic Publishers, 1996.

[5] L.Kleinrock and F. Tobagi, "Packet switch in radio channels: Part I-carrier sense multiple-access modes and their throughput-delay characteristics ," *IEEE Trans. Commun.*, vol. COM-23, pp. 1400–1416, Dec. 1975.

---

[1] The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

EFFICIENT SORTING ON THE STAR GRAPH INTERCONNECTION NETWORK

Selim G. Akl and Tanya Wolff
Department of Computing and Information Science
Queen's University, Kingston, Ontario, Canada K7L 3N6

Abstract

Two algorithms for sorting $n!$ numbers on an n-star interconnection network are described. Both algorithms are based on arranging the $n!$ processors of the n-star in a virtual $(n-1)$-dimensional array. The first algorithm runs in $O(n^3 \log n)$ time. This performance matches that of the fastest previously known algorithm for the same problem. The second algorithm runs in $O(n^2)$ time and thus provides an asymptotic improvement over its predecessors.

1 Introduction

For a positive integer $n > 1$, an *n-star interconnection network*, denoted $\mathcal{S}_n$, is defined as follows:

1. $\mathcal{S}_n$ is an undirected graph with $n!$ vertices, each of which is a processor;
2. The label v of each processor P_v is a distinct permutation of the symbols $\{1, 2, \ldots, n\}$;
3. Processor P_v is directly connected by an edge to each of $n-1$ processors P_u, where u is obtained by interchanging the first and ith symbols of v, that is, if

$$v = v(1)v(2)\ldots v(n),$$

where $v(j) \in \{1, 2, \ldots, n\}$ for $j = 1, 2, \ldots, n$, then

$$u = v(i)v(2)\ldots v(i-1)v(1)v(i+1)\ldots v(n),$$

for $i = 2, 3, \ldots, n$.

Because its degree and diameter are both sublogarithmic in the number of vertices [1], the n-star interconnection network has received a fair bit of attention lately [3], [5], [7], [8], [12]. Of particular interest in this paper is the problem of sorting on the n-star. Specifically, given $n!$ numbers, held one per processor of the n-star, it is required to sort these numbers in nondecreasing order. The previous best algorithm for sorting $n!$ numbers on an n-star has a running time of $O(n^3 \log n)$ [12].

In this paper we describe two algorithms for sorting on $\mathcal{S}_n$. The central idea in both algorithms is to map the processors of the n-star onto an $(n-1)$-dimensional array. The first algorithm uses a generalization of the algorithm in [14] and runs in $O(n^3 \log n)$ time.

This performance matches that of the algorithm in [12]. However, the new algorithm is considerably simpler to state and provides an alternative paradigm for sorting on the n-star. Furthermore, it can be expressed in a concise and straightforward manner without the need for recursion and, therefore, may be easier (and perhaps more efficient) to implement. The second algorithm is directly based on an algorithm developed in [10] for sorting on an $(n-1)$-dimensional lattice. It runs in $O(n^2)$ time, thus achieving an asymptotic improvement over the fastest previous algorithm [12].

2 Ordering the Processors

As mentioned in Section 1, each of the $n!$ processors of the n-star holds one of the $n!$ numbers to be sorted. To properly define the problem of sorting, it is imperative to impose an order on the processors. This way, one can say that the set of numbers has been sorted once the smallest number of the set resides in the 'first' processor, the second smallest number of the set resides in the 'second' processor, and so on. (If two numbers are equal then the one with the smaller initial index precedes the other). One such order on the processors is now defined.

For a given permutation $v = v(1)v(2)\ldots v(n)$, let

(a) $vmax(n) = \max\{v(k) \mid v(k) < v(n),\ 1 \le k \le n-1\}$, if such a $v(k)$ exists; otherwise, $vmax(n) = \max\{v(k) \mid 1 \le k \le n-1\}$.

(b) $vmin(n) = \min\{v(k) \mid v(k) > v(n),\ 1 \le k \le n-1\}$, if such a $v(k)$ exists; otherwise, $vmin(n) = \min\{v(k) \mid 1 \le k \le n-1\}$.

Beginning with the processor label $v = 12\ldots n$, the remaining $n! - 1$ labels are arranged in the order produced by algorithm LABELS (n, j) given below. The algorithm operates throughout on the array $v = v(1)v(2)\ldots v(n)$. Initially, $v(i) = i$, $1 \le i \le n$, and the algorithm is called with $j = 1$. A new permutation is produced every time two elements of the array v are swapped.

```
Algorithm LABELS($n, j$)
  Step 1: $i \leftarrow 1$
  Step 2: while $i \le n$ do
            (2.1) if $n > 2$
                    then LABELS($n-1, i$)
                  end if
            (2.2) if $i < n$
                    then if $j$ is odd
                           then $v(n) \leftrightarrow vmax(n)$
                           else $v(n) \leftrightarrow vmin(n)$
                         end if
                  end if
            (2.3) $i \leftarrow i + 1$
          end while. □
```

The order produced by the above algorithm satisfies the following properties:

1. All $(n-1)!$ permutations with $v(n) = \ell$ precede all permutations with $v(n) = \ell'$, for each ℓ and ℓ' such that $1 \le \ell' < \ell \le n$.

2. Consider the jth group of $(m-1)!$ consecutive permutations in which $v(m)v(m+1)\ldots v(n)$ are fixed, where $1 \leq j \leq n!/(m-1)!$ and $2 \leq m \leq n-1$. Within this group, the $m-1$ elements of the set

$$\{1, 2, \ldots, n\} - \{v(m), v(m+1), \ldots, v(n)\}$$

take turns in appearing as $v(m-1)$, each repeated $(m-2)!$ times consecutively. They appear in decreasing order if j is odd and in increasing order if j is even.

For example, for $n = 4$, the 4! labels, that is, the 24 permutations of $\{1, 2, 3, 4\}$, ordered according to algorithm LABELS, are as follows:

1. 1234
2. 2134
3. 3124
4. 1324
5. 2314
6. 3214
7. 4213
8. 2413
9. 1423
10. 4123
11. 2143
12. 1243
13. 1342
14. 3142
15. 4132
16. 1432
17. 3412
18. 4312
19. 4321
20. 3421
21. 2431
22. 4231
23. 3241
24. 2341

3 Rearranging the Data

An important operation in our sorting algorithms is data rearrangement whereby pairs of processors exchange their data. We now define this operation. Recall that in any interconnection network, two processors that are directly connected by an edge are called *neighbors*. A processor can send a datum of constant size to a neighbor in constant time. In $\mathcal{S}_n$ each processor P_v has $n-1$ neighbors P_u. The label u of the kth neighbor of P_v is obtained by swapping $v(1)$ with $v(k+1)$, for $k = 1, 2, \ldots, n-1$. The edges connecting P_v to its neighbors are referred to as *connections* $2, 3, \ldots, n$. If each processor of $\mathcal{S}_n$ holds a datum, then the phrase "rearranging the data over connection i" means that all processors directly connected through connection i exchange their data. Specifically, each processor P_v, where $v = v(1)v(2)\ldots v(i-1)v(i)v(i+1)\ldots v(n)$ sends its datum to processor P_u, where $u = v(i)v(2)\ldots v(i-1)v(1)v(i+1)\ldots v(n)$.

4 The Star Viewed As a Multidimensional Array

The $n!$ processors of $\mathcal{S}_n$ are thought of as being organized in a virtual $(n-1)$-dimensional $2 \times 3 \times \cdots \times n$ orthogonal array. This organization is achieved in two steps: The processors are first listed according to the order obtained from algorithm LABELS, they are then placed (figuratively) in this order in the $(n-1)$-dimensional array. For example, for $n = 4$, the 24 processor labels appear in the $2 \times 3 \times 4$ array as shown below:

$$\begin{array}{cc@{\qquad}cc@{\qquad}cc@{\qquad}cc}
1234 & 2134 & 1243 & 2143 & 1342 & 3142 & 2341 & 3241 \\
1324 & 3124 & 1423 & 4123 & 1432 & 4132 & 2431 & 4231 \\
2314 & 3214 & 2413 & 4213 & 3412 & 4312 & 3421 & 4321
\end{array}$$

It is important to note that the processors are placed in *snakelike order by dimension*. When $n = 2$, this corresponds to the familiar *snakelike row-major order*. The rule for placing the processors is as follows. Each position of the $(n-1)$-dimensional array has coordinates $I(1), I(2), \ldots, I(n-1)$, where $1 \leq I(k) \leq k+1$. For example, for $n = 4$, the coordinates of the 24 positions in the $2 \times 3 \times 4$ array are as follows:

$1,1,1 \quad 2,1,1 \qquad 1,1,2 \quad 2,1,2 \qquad 1,1,3 \quad 2,1,3 \qquad 1,1,4 \quad 2,1,4$
$1,2,1 \quad 2,2,1 \qquad 1,2,2 \quad 2,2,2 \qquad 1,2,3 \quad 2,2,3 \qquad 1,2,4 \quad 2,2,4$
$1,3,1 \quad 2,3,1 \qquad 1,3,2 \quad 2,3,2 \qquad 1,3,3 \quad 2,3,3 \qquad 1,3,4 \quad 2,3,4$

In dimension k, $1 \leq k \leq n-1$, there are $2 \times 3 \times \cdots \times k \times (k+2) \times \cdots \times n$ (i.e., $n!/(k+1)$) groups of $k+1$ consecutive positions. The positions in each group have the same

$$I(1), I(2), \ldots, I(k-1), I(k+1), I(k+2), \ldots, I(n-1)$$

coordinates (in other words, they differ only in coordinate $I(k)$, with the first position in the group having $I(k) = 1$, the second $I(k) = 2$, and the last $I(k) = k+1$).

The following recursive function $snake_k$ [10] maps the coordinates of each position of a multidimensional array into an integer: $snake_1(I(1)) = I(1)$, for $I(1) = 1, 2$, and

$$\begin{aligned} snake_k(I(1), I(2), \ldots, I(k)) &= k!(I(k)-1) \\ &+ \begin{cases} snake_{k-1}(I(1), I(2), \ldots, I(k-1)) & \text{if } I(k) \text{ is odd} \\ k! + 1 - snake_{k-1}(I(1), I(2), \ldots, I(k-1)) & \text{if } I(k) \text{ is even,} \end{cases} \end{aligned}$$

for $2 \leq k \leq n-1$, $1 \leq I(i) \leq i+1$, and $1 \leq i \leq k$. Thus the function $snake_k$ maps the coordinates $I(1), I(2), \ldots, I(n-1)$ of each position of the $(n-1)$-dimensional $2 \times 3 \times \cdots \times n$ array into a unique integer in the set $\{1, 2, \ldots, n!\}$ according to snakelike order by dimension.

As a result of the (figurative) arrangement just described, processor P_v, such that $v = v(1)v(2)\ldots v(n)$, occupies that position of the $(n-1)$-dimensional array whose coordinates $I(1), I(2), \ldots, I(n-1)$ are given by

$$I(k) = k + 1 - \sum_{j=1}^{k} [v(k+1) > v(j)],$$

for $1 \leq k \leq n-1$, where $[v(k+1) > v(j)]$ equals 1 if $v(k+1) > v(j)$ and equals 0 otherwise.

It should be stressed here that two processors occupying adjacent positions in some dimension k, $1 \leq k \leq n-1$, on the $(n-1)$-dimensional array are not necessarily directly connected on the n-star. Now suppose that in dimension k, each processor in a group of $k+1$ processors occupying consecutive positions holds a datum. Our purpose in what follows is to show that after rearranging the data over connection $k+1$, these $k+1$ data are stored in $k+1$ processors forming a linear array (i.e., the first of the $k+1$ processors is directly connected to the second, the second is directly connected to the third, and so on). To simplify the presentation, we assume in what follows that the $k+1$ consecutive positions (of the multidimensional array) occupied by the $k+1$ processors begin at the position with $I(k) = 1$ and end at that with $I(k) = k+1$ (the argument being symmetric for the case where the $k+1$ processors are placed in the opposite direction, that is, from the position with $I(k) = k+1$ to that with $I(k) = 1$). We proceed in two steps.

1. First we show how the labels of these $k+1$ processors, which are permutations of $\{1, 2, \ldots, n\}$, can be obtained from one another. Let $v(1)v(2)\ldots v(k+1)\ldots v(n)$ be the label of the last processor in the group, and let d_m, $1 \leq m \leq k+1$, be the position of the mth smallest symbol among $v(1)v(2)\ldots v(k+1)$, where $d_1 = k+1$. Then, the label of the $(k+1-i)$th processor in the group, $1 \leq i \leq k$, is obtained from the label of the $(k+1-i+1)$st processor by exchanging the symbols in

positions $k+1$ and d_{i+1}. For example, for $n = 4$, consider the four processors, occupying consecutive positions over dimension 3, whose labels are 1324, 1423, 1432, and 2431. The second of these (i.e., 1423) can be obtained from the third (i.e., 1432) by exchanging the symbol in position 4 (i.e., 2) with the third smallest symbol among 2, 4, 3, and 1 (i.e., 3). This property follows directly from the definition of the LABEL ordering.

2. Now consider a group of $k+1$ processors occupying consecutive positions over dimension k, with each processor holding a datum and the label of the $(k+1)$st processor in the group being $v(1)v(2)\ldots v(k+1)\ldots v(n)$. From the property established in the previous paragraph, the label of the $(k+1-i)$th processor has the $(i+1)$st smallest symbol among $v(1)v(2)\ldots v(k+1)$ in position $k+1$, for $0 \leq i \leq k$. Rearranging the data over connection $k+1$ moves the data from this group to another group of $k+1$ processors in which the label of the $(k+1-i)$th processor has the $(i+1)$st smallest symbol among $v(1)v(2)\ldots v(k+1)$ in position 1, for $0 \leq i \leq k$. In this new group, the label of the $(k+1-i)$th processor can be obtained from that of the $(k+1-i+1)$st by exchanging the symbol in position 1 with that in position d_{i+1}, for $1 \leq i \leq k$. It follows that the processors in the new group are connected to form a linear array. For example, consider once again the four processors, occupying consecutive positions over dimension 3, whose labels are 1324, 1423, 1432, and 2431. Rearranging the data held by these processors over connection 4 moves them to the processors whose labels are 4321, 3421, 2431, and 1432, respectively.

5 A Simple Sorting Algorithm for $\mathcal{S}_n$

Let each of the $n!$ processors of the n-star hold a number. It is required to sort these numbers in nondecreasing order. We now present an algorithm that solves this problem by viewing the n-star as an $(n-1)$-dimensional array. The $n!$ numbers are sorted on the array in snakelike order by dimension, as defined in Section 4. This order matches that of the processors on the $(n-1)$-dimensional array.

The algorithm consists of a number of iterations. During an iteration, the numbers are sorted either in the forward or reverse direction along each of the $n-1$ dimensions in turn. Here, "sorting in the forward (reverse) direction in dimension k" means sorting the numbers in nondecreasing (nonincreasing) order from the processor with $I(k) = 1$ to the processor with $I(k) = k+1$. This sorting is performed using an algorithm for sorting $k+1$ numbers on a linear array of $k+1$ processors in $k+1$ steps, that is, $O(k)$ time, described in [3]. Note that since in dimension k the $k+1$ processors in each group are not necessarily connected on the n-star to form a linear array, their contents are copied in constant time to another set of $k+1$ processors which *are* connected on the n-star as a linear array. This is done simply by rearranging the numbers to be sorted over connection $k+1$, as described in Section 3. After the sorting is performed, the numbers are brought back to the original processors, also in constant time. Henceforth, we refer to each group of consecutive positions in a given dimension of the multidimensional array, as a 'row'. Thus, a row in dimension k consists of $k+1$ consecutive positions. Let $D(k) = \sum_{r=k+1}^{n-1}(I(r)-1)$, for $k = 1, 2, \ldots, n-1$, where the empty sum (when $k = n-1$) is equal to 0 by definition, and let $N = \sum_{k=2}^{n}\lceil \log k \rceil$.
The algorithm is as follows:

```
**Algorithm SORTING ON STAR**
  for $i = 1$ to $N$ do
    for $k = 1$ to $n - 1$ do
      for each row in dimension $k$ do in parallel
        if $D(k)$ is even
        then sort in the forward direction
        else sort in the reverse direction
        end if
      end for
    end for
  end for. □
```

For example, let $n = 5$, and let the set to be sorted be $\{1, 2, \ldots, 120\}$. The numbers are initially stored in the 120 processors of S_5, one number per processor, in arbitrary order. Once algorithm SORTING ON STAR has been applied, the following arrangement results:

1	2	12	11	13	14	24	23
4	3	9	10	16	15	21	22
5	6	8	7	17	18	20	19
48	47	37	38	36	35	25	26
45	46	40	39	33	34	28	27
44	43	41	42	32	31	29	30
49	50	60	59	61	62	72	71
52	51	57	58	64	63	69	70
53	54	56	55	65	66	68	67
96	95	85	86	84	83	73	74
93	94	88	87	81	82	76	75
92	91	89	90	80	79	77	78
97	98	108	107	109	110	120	119
100	99	105	106	112	111	117	118
101	102	104	103	113	114	116	115

We note here, as pointed out earlier, that algorithm SORTING ON STAR is nonrecursive, making it straightforward to implement. Its idea of repeatedly going through the dimensions of the multidimensional array during each iteration of the outer **for** loop is reminiscent of the ASCEND paradigm (originally proposed for the hypercube and related interconnection networks [3]).

Analysis. We now show that N iterations, each of which sorts the rows in dimensions $1, 2, \ldots, n - 1$, suffice to correctly sort $n!$ numbers stored in an $(n - 1)$-dimensional $2 \times 3 \times \cdots \times n$ array. First note that the algorithm for sorting on a linear array given in [3] is *oblivious*, that is, the sequence of comparisons it performs is predetermined. It follows that algorithm SORTING ON STAR is also oblivious. This property allows us to use the 0-1 principle [3]: If we can show that the algorithm correctly sorts any sequence of 0s and 1s, it will follow that the algorithm correctly sorts any sequence of numbers. Suppose then that the input to algorithm SORTING ON STAR consists of an arbitrary sequence of 0s and 1s. Once sorted in nondecreasing order, this sequence will consist of a (possibly empty) subsequence of 0s followed by a (possibly empty) subsequence of 1s. In particular, the sorted sequence will contain at most one (0, 1) pattern, that is, a 0 followed by a 1. (If the input consists of all 0s or all 1s, then of course the output contains no such (0, 1) pattern.) If present, the (0, 1) pattern will appear either in a row in dimension 1, or at the boundary between two adjacent dimension 1 rows (consecutive in snakelike order) such that the 0 appears in one row and the 1 in the next.

Now consider the rows in dimension k, $1 \le k \le n-1$. There are $n!/(k+1)$ such rows. A row in dimension k is said to be *clean* if it holds all 0s (all 1s) and there are no 1s preceding it (no 0s following it) in snakelike order; otherwise, the row is *dirty*.

Suppose that an iteration of the algorithm has just sorted the contents of the rows in dimension $n-2$, and let us focus on two such rows, adjacent in dimension $n-1$. For example, for $n = 4$, the coordinates of two dimension 2 rows, adjacent in dimension 3 of the $2 \times 3 \times 4$ array, are

$$\begin{matrix} 1,1,1 \\ 1,2,1 \\ 1,3,1 \end{matrix} \quad \text{and} \quad \begin{matrix} 1,1,2 \\ 1,2,2 \\ 1,3,2, \end{matrix}$$

respectively. Because these two dimension $n-2$ rows were sorted in snakelike order, they yield at least one clean dimension $n-2$ row when the dimension $n-1$ rows are sorted. To illustrate, suppose that the two dimension 2 rows above contained

$$\begin{matrix} 0 \\ 0 \\ 1 \end{matrix} \quad \text{and} \quad \begin{matrix} 1 \\ 0 \\ 0, \end{matrix}$$

respectively after they were sorted. Then (regardless of what the other dimension 2 rows contained) a clean dimension 2 row containing all 0s is created after the dimension 3 rows are sorted. Thus, after $\lceil \log n \rceil$ iterations all dimension $n-2$ rows are clean, except possibly for $(n-2)!$ dimension $n-2$ rows contiguous in dimension $n-2$. All dimension $n-1$ rows are now permanently sorted.

By the same argument, $\lceil \log(n-1) \rceil$ iterations are subsequently needed to make all dimension $n-3$ rows clean (except possibly for $(n-3)!$ dimension $n-3$ rows contiguous in dimension $n-3$). In general, $\lceil \log(k+1) \rceil$ iterations are needed to make all dimension $k-1$ rows clean (except possibly for $(k-1)!$ dimension $k-1$ rows contiguous in dimension $k-1$) after all dimension k rows have been permanently sorted, for $k = n-1, n-2, \ldots, 2$. This leaves possibly one dirty dimension 1 row holding the (0, 1) pattern. Since that row may not be sorted in the proper direction for snakelike ordering, one final iteration completes the sort. The algorithm therefore sorts correctly in $N = \lceil \log n \rceil + \lceil \log(n-1) \rceil + \cdots + \lceil \log 3 \rceil + 1$, that is, $O(n \log n)$ iterations.

Finally, observe that the algorithm of [3] for sorting on a linear array is used in dimensions $1, 2, \ldots, n-1$, and hence each iteration of the outer **for** loop in algorithm SORTING ON STAR requires $\sum_{k=1}^{n-1} O(k) = O(n^2)$ time. Since the outer loop is executed $O(n \log n)$ times, the algorithm has a running time of $O(n^3 \log n)$.

The previous discussion suggests that the running time of algorithm SORTING ON STAR may in fact be reduced by a constant factor as follows. Because all dimension

$n-1$ rows are permanently sorted after $\lceil \log n \rceil$ iterations of the outer **for** loop, there is no purpose in applying any sorting operations to these rows in subsequent iterations. After another $\lceil \log(n-1) \rceil$ iterations, dimension $n-2$ rows require no more attention. In general, after $\sum_{k=j+1}^{n} \lceil \log k \rceil$ iterations, dimension j rows need not be sorted.

6 An Efficient Sorting Algorithm for $\mathcal{S}_n$

Suppose that M processors, where $M = n_1 n_2 \cdots n_r$ and each n_i is a positive integer, reside in an r-dimensional $n_1 \times n_2 \times \cdots \times n_r$ array such that the processors in each row in dimension k, $1 \leq k \leq r$, are connected to form a linear array. This interconnection network is known as an *r-dimensional lattice* (or *grid*, or *mesh*). Clearly, the special case where $M = n!$, $r = n-1$, and $n_i = i+1$, $1 \leq i \leq n-1$, is of particular interest in this paper, and we refer to the $(n-1)$-dimensional $2 \times 3 \times \cdots \times n$ network as an *$n!$-lattice*. The *diameter* of the $n!$-lattice, that is, the shortest distance separating the farthest two processors, is $1 + 2 + \cdots + n - 1 = n(n-1)/2$. It follows therefore that a lower bound on the number of steps required to sort $n!$ numbers on the $n!$-lattice is $\Omega(n^2)$.

Algorithms for the problem of sorting $n!$ numbers on the $n!$-lattice can be derived from [4], [6], [9], [10], [13], and [15]. The deterministic and oblivious algorithm of Section 5 can also be used to sort $n!$ numbers on the $n!$-lattice. However, many of these algoritms have running times that exceed n^2 asymptotically and, as a result, are not (worst-case) optimal in light of the $\Omega(n^2)$ lower bound. By contrast, the (deterministic and oblivious) algorithm obtained from [10] runs in $O(n^2)$ time and is therefore optimal in the worst case. Furthermore, like algorithm SORTING ON STAR, the algorithm of [10] arranges the processors in snakelike order by dimension.

In what follows we provide a brief description of how the algorithm of [10] sorts $M = n_1 n_2 \cdots n_r$ data stored in an r-dimensional lattice. For each n_i, let u_i and v_i be positive integers such that $u_i v_i = n_i$, $1 \leq i \leq r$. The r-dimensional lattice is viewed as consisting of $u_1 u_2 \cdots u_r$ r-dimensional $v_1 \times v_2 \times \cdots \times v_r$ sublattices, called *blocks*, numbered in snakelike order by dimension from 1 to $u_1 u_2 \cdots u_r$. Alternatively, the r-dimensional lattice consists of $u_1 u_2 \cdots u_r$ $(r-1)$-dimensional *hyperplanes* of $v_1 v_2 \cdots v_r$ processors each, numbered from 1 to $u_1 u_2 \cdots u_r$. The algorithm is given below.

```
Algorithm SORTING ON LATTICE
  Step 1: Sort the contents of block $i$, $1 \leq i \leq u_1 u_2 \cdots u_r$.
  Step 2: Move the data held by the processors in block $i$
          to the corresponding processors in hyperplane $i$, $1 \leq i \leq u_1 u_2 \cdots u_r$.
  Step 3: Sort the contents of all $r$-dimensional
          $v_1 \times v_2 \times \cdots \times n_r$ 'towers' of blocks.
  Step 4: Move the data held by the processors in hyperplane $i$
          to the corresponding processors in block $i$, $1 \leq i \leq u_1 u_2 \cdots u_r$.
  Step 5: Sort the contents of
          (5.1) All pairs of consecutive blocks $i$ and $i+1$,
                for all odd $i$, $1 \leq i \leq u_1 u_2 \cdots u_r$.
          (5.2) All pairs of consecutive blocks $i-1$ and $i$,
                for all even $i$, $2 \leq i \leq u_1 u_2 \cdots u_r$. □
```

The algorithm requires $O(n_1 + n_2 + \cdots + n_r)$ elementary steps (in which a datum is sent from a processor to its neighbor). For an $n!$-lattice (i.e., when $r = n-1$ and $n_i = i+1$, $1 \leq i \leq n-1$) the running time is $O(n^2)$.

The $n!$-lattice is essentially the $(n-1)$-dimensional array of Section 4, augmented with edges connecting the processors. Therefore, any constant-time exchange of data between two neighboring processors on the $n!$-lattice can be executed on the n-star in constant time. Consequently, using the techniques described in Sections 2 - 4, algorithm SORTING ON LATTICE directly translates into an algorithm for sorting on $\mathcal{S}_n$ in $O(n^2)$ time.

7 Conclusion

Two algorithms for sorting $n!$ numbers on an n-star interconnection network were described. Both algorithms are based on arranging the $n!$ processors of $\mathcal{S}_n$ in a virtual $(n-1)$-dimensional array. The first algorithm sorts the input by using the simple idea of repeatedly sorting all the rows in each dimension. Its running time is $O(n^3 \log n)$. While our main objective was to develop an algorithm for sorting on the n-star, the algorithm of Section 5 is effectively an algorithm for sorting on an $n!$-lattice and is sufficiently interesting in its own right. It is nonrecursive and does not require that data be exchanged between pairs of distant processors: The only operation used, namely, 'comparison-exchange', applies to neighboring processors. Thus, because no routing is needed, the algorithm has virtually no control overhead. Note also that this algorithm, while intended in Section 5 for a $2 \times 3 \times \cdots \times n$ array, is easily extended to arbitrary $n_1 \times n_2 \times \cdots \times n_r$ r-dimensional lattices. In particular, when $n_i = 2$ for $1 \leq i \leq r$, a very simple algorithm is obtained for sorting $M = 2^r$ numbers on a hypercube with M processors in $O(\log^2 M)$ time. Other algorithms for sorting on the hypercube, including those with the same running time (i.e., $O(\log^2 M)$) as well as asymptotically faster ones, are considerably more complicated [2, 11].

The second algorithm described in this paper for sorting on $\mathcal{S}_n$ is essentially an adaptation of an algorithm appearing in the literature for sorting on a multidimensional lattice of processors. It runs in $O(n^2)$ time. In view of the $\Omega(n^2)$ lower bound on the running time of any algorithm for sorting on the lattice, this performance is the best that one can hope for using the approach taken in this paper. Nonetheless, this represents an improvement by a factor of $O(n \log n)$ over the best previous algorithm for sorting on $\mathcal{S}_n$.

Since $n!$ numbers are sorted optimally using one processor in $O(n! \log n!)$ time, a parallel algorithm using $n!$ processors must have a running time of $O(\log n!)$, that is, $O(n \log n)$, in order to be time-optimal. In that sense, the algorithm of Section 6 is suboptimal by a factor of $O(n/\log n)$. However, it is not known at the time of this writing whether an $O(n \log n)$ running time for sorting $n!$ numbers is achievable on the n-star.

8 Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada. The authors wish to thank Christian Lavault of the University of Paris for interesting discussions.

References

[1] S.B. Akers, D. Harel, and B. Krishnamurthy. The star graph: An attractive al-

ternative to the n-cube. *Proceedings of the International Conference on Parallel Processing*, 1987, pp. 393–400.

[2] S.G. Akl. *Parallel Sorting Algorithms.* Academic Press, Orlando, Florida, 1985.

[3] S.G. Akl. *Parallel Computation: Models and Methods.* Prentice Hall, Upper Saddle River, New Jersey, 1997.

[4] K. Brockmann and R. Wanka. Efficient oblivious parallel sorting on the MasPar MP-1. *Proceedings of the Hawaii Conference on System Sciences*, Vol. 1, 1997, pp. 200–208.

[5] W.K. Chiang and R.J. Chen. The (n, k)-star graph: A generalized star graph. *Information Processing Letters*, Vol. 56, 1995, pp. 259–264.

[6] P.F. Corbett and I.D. Scherson. A unified algorithm for sorting on multidimensional mesh-connected processors. *Information Processing Letters*, Vol. 37, 1991, pp. 225–231.

[7] M. Dietzfelbinger, S. Madhavapeddy, and I.H. Sudborough. Three disjoint path paradigms in star networks. *Proceedings of the IEEE Symposium on Parallel and Distributed Processing*, 1991, pp. 400–406.

[8] J.S. Jwo, S. Lakshmivarahan, and S.K. Dhall. Embedding of cycles and grids in star graphs. *Proceedings of the IEEE Symposium on Parallel and Distributed Processing*, 1990, pp. 540–547.

[9] M. Kaufmann, J.F. Sibeyn, and T. Suel. Beyond the worst-case bisection bound: Fast sorting and routing on meshes. *Proceedings of the European Symposium on Algorithms*, Lecture Notes in Computer Science No. 979, Springer-Verlag, Berlin, 1995, pp. 75–88.

[10] M. Kunde. Optimal sorting on multi-dimensionally mesh-connected computers. *Proceedings of the Symposium on Theoretical Aspects of Computer Science*, Lecture Notes in Computer Science No. 247, Springer-Verlag, Berlin, 1987, pp. 408–419.

[11] F.T. Leighton. *Introduction to Parallel Algorithms and Architectures.* Morgan Kaufmann, San Mateo, California, 1992.

[12] A. Menn and A.K. Somani. An efficient sorting algorithm for the star graph interconnection network. *Proceedings of the International Conference on Parallel Processing*, Vol. 3, 1990, pp. 1–8.

[13] D. Nassimi and S. Sahni. Bitonic sort on a mesh-connected parallel computer. *IEEE Transactions on Computers*, Vol. C-27, 1979, pp. 2–7.

[14] I. Scherson, S. Sen, and A. Shamir. Shear-sort: A true two-dimensional sorting technique for VLSI networks. *Proceedings of the International Conference on Parallel Processing*, 1986, pp. 903–908.

[15] C.D. Thompson and H.T. Kung. Sorting on a mesh-connected parallel computer. *Communications of the ACM*, Vol. 20, 1977, pp. 263–271.

Matching Output Queueing with Combined Input and Output Queueing

Nick McKeown*, Balaji Prabhakar‡, and Mingyan Zhu*

*Department of Electrical Engineering, Stanford University, Stanford, CA 94305-9030, USA.
‡Hewlett-Packard BRIMS, Filton Road, Stoke Gifford, Bristol BS12 6QZ, England.

Abstract

At very high aggregate bandwidths, output queueing is impractical because of insufficient memory bandwidth. This problem is getting worse: memory bandwidth is improving slowly, whereas the demand for network bandwidth continues to grow exponentially. The difficulty is that output-queued switches require memories that run at a speedup of N, where N is equal to the number of switch ports. This paper addresses the following question: Is it possible for a switch to ***exactly*** *match output-queueing with a reduced speedup? We prove that if virtual output queueing is used, a combined input-output queued switch is always work-conserving if its speedup is greater than $N/2$. This result is proved using a novel scheduling algorithm - the Home Territory Algorithm (HTA).*

1: Introduction

Many commercial switches and routers today employ output queueing.[1] The advantages of output queueing are two-fold. First, it allows the throughput to be maximized: so long as no input or output is oversubscribed, the switch will support the traffic. Second, because packets are immediately placed in output queues upon arrival, it is possible to control the latency of packets through the switch. This is very important for supporting QoS in a switch or router. But output queueing is complex and expensive. Its complexity arises from a need for the switch fabric and memory to run N times as fast as the line rate, assuming an $N \times N$ switch. The ratio of a switch's internal bandwidth to its line rate is defined as the speedup of a switch. Hence, an output-queued (OQ) switch has a speedup of N.

Input queueing has lower complexity, and consequently lower cost. The switch fabric and the memory at the inputs of an input-queued (IQ) switch need only run as fast as the line rate. In other words, an input-queued switch has a speedup of 1. However, if each input maintains a single FIFO queue, input queueing suffers head of line (HOL) blocking, which limits the throughput to just 58.6%. For non-uniform traffic, the performance is even worse [1].

Output queueing is limited by the bandwidth of commercially available memories. Currently, it seems practical to implement an output-queued switch or router with an aggregate bandwidth of approximately 20Gb/s. But the continued exponential increase in demand for bandwidth is making faster and faster switches necessary. Before long, it may be impractical to build output-queued switches.

A lot of research has been done to approach the performance of output queueing using a combined input-output queued switch. A combined input-output queued (CIOQ) switch is defined as a non-blocking packet switch with N inputs and N outputs and operates S times as fast as the line rate. With a speedup of S, a CIOQ switch can remove up to S packets from each input and deliver up to S packets to each output within a time slot. A time slot is the time between packet arrivals at input ports. Since $1 < S < N$, packets need to be buffered before switching at the inputs as well as after switching at the outputs.

---

[1] When we refer to output queueing in this paper, we include designs that employ centralized shared memory.

Various properties of CIOQ switches have been studied, such as average packet delay, maximum throughput, packet loss probability, packet blocking probability, optimal buffer allocation, effects of arbitration policy on throughput, etc.[2][3][4] and [5]. It has been shown that, when the input traffic is independent and uniform, a CIOQ switch can achieve 99% throughput with a modest speedup of approximately four [3][4][5].

The aim of our work is different. Rather than find values of speedup that work well on average, or with relatively benign types of traffic, we aim to find the minimum speedup such that a CIOQ switch behaves *identically* to an OQ switch for *all* types of traffic. Here, 'behave identically' means that by observing only the output processes, an OQ switch is indistinguishable from a CIOQ switch. The output processes are indistinguishable if and only if they are busy at the same time, and have the same packet departure order.

An OQ switch is work-conserving since its outputs never idle when there are packets in the system for them. Therefore, a CIOQ switch that operates identically, is also work-conserving. In our work, we have searched for techniques such that a CIOQ switch can be work-conserving. Note, a work-conserving CIOQ switch may not have the same packet departure order as an OQ switch.

Our work has two results. First, we formalized the intuitive notion that a $N \times N$ switch with a single FIFO queue at each input requires speedup $S = N$ to be work conserving; i.e. it must be an output-queued switch. Second, we found that a $N \times N$ CIOQ switch that uses *virtual output queueing*[1] can be work conserving if $S > \frac{N}{2}$, *for any type of input traffic*. We proved this result using a novel scheduling algorithm: the Home Territory Algorithm (HTA).

This paper is organized as follows: Section 2 describes our combined input-output queued packet switch model, Section 3 discusses the main results of our research, and Section 4 summarizes our findings.

2: Our CIOQ Packet Switch Model

The combined input-output queued switch we consider, shown in Figure 1, is a single-stage packet switch. It has N inputs, N outputs, a non-blocking space division switch fabric, and a scheduler. Both the inputs and the outputs are labelled from 0 to $N-1$.

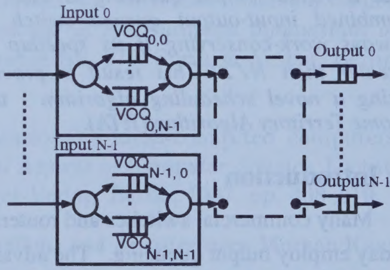


Figure 1: Our CIOQ Packet Switch

The switch we consider operates in synchronous mode where packets arrive and are transmitted synchronously. Packets are assumed to have fixed length and fixed transmission time. We choose the time between packet arrivals as our time unit, and call it a time slot.

---

[1] This is a technique commonly used in input-queued switches to eliminate HOL blocking. With virtual output queueing, a switch maintains a separate FIFO at each input for each output.

The switch has a speedup of S, i.e. its fabric runs S times as fast as the line rate, where $1 < S < N$. A $N \times N$ switch with speedup of S can remove at most S packets from each input and can transmit at most S packets to each output in a time slot.

A CIOQ switch requires buffering at its inputs as well as at its outputs. We assume that all buffers have infinite capacity. Each output maintains a FIFO queue, and operates on a first come first serve basis. Each input maintains a separate FIFO queue for each output. Hence, there are N FIFO queues at each input. These FIFO queues are Virtual Output Queues(VOQ), where VOQ_{ij} denotes a FIFO queue at input i that queues packets destined to output j.

A scheduling algorithm selects a matching between the inputs and outputs. A matching is selected in such a way that each non-empty input is connected to at most one output and each output is connected to at most one input. For a CIOQ switch with a speedup of S, a scheduling algorithm carries out the matching S times per time slot. Our switch uses a novel scheduling algorithm called the Home Territory Algorithm (HTA). Section 3 gives a detailed description of HTA.

3: Main Results

3.1: Without Virtual Output Queues

It is intuitively clear that if a CIOQ switch maintains a single FIFO queue at each input, then a speedup of N is sufficient for it to be work-conserving. In fact, it is both necessary and sufficient:

Theorem 1: *For a CIOQ switch that maintains a single FIFO queue at each input, a speedup of N is required for it to be work-conserving.*

Proof: See Appendix A.

In an effort to match the performance of output-queueing, it is worth asking the questions: Is this limited by HOL blocking? If not, can we arrange the input queues differently and use a lower speedup to make the switch conserve work? As we shall see, the answer to the first question is "No!", whereas the answer to the second is "Yes!".

3.2: With Virtual Output Queueing.

In this section, we describe a novel scheduling algorithm - the Home Territory Algorithm - and prove our main result: A CIOQ switch that uses VOQ buffering and operates under HTA conserves work for all types of input traffic when its speedup is greater than $N/2$.

3.2.1: Definitions

Definition: *Phases.*

A time slot is divided into S equal intervals, where S is the speedup of the switch. Each of these intervals is called a *phase.* During phase i, denoted Φ_i, $1 \le i \le S$, a switch can remove at most one packet from each input, and can transfer at most one packet to each output.

Definition: *Output Home Territory (HT^O) and Non Home Territory (NHT^O) sets.*

Each output partitions the set of inputs into two disjoint subsets: its Home Territory (HT^O) set and its Non Home Territory (NHT^O) set. The HT^O set of output j is denoted $HT^O(j)$, and similarly the NHT^O set of output j, $NHT^O(j)$. This partitioning is based on the following static priority that each output assigns to each input.

The priority assigned by output j to input i is denoted $P_j^O(i)$. If $i > j$, then $P_j^O(i) = i - j$, else $P_j^O(i) = (i + N) - j$. In other words, output j assigns the highest priority to input $(j + 1)$ Mod N, the second highest priority to input $(j + 2)$ Mod N, and so on.

Every output ranks all inputs from priority 1 to N. Here, smaller number indicates higher priority. The top S inputs constitute its NHT^O set and the remaining inputs constitute its HT^O set, so that $|NHT^O| = S$ and $|HT^O| = N - S$.

Table 1 gives the territory sets of output j and lists the priorities it assigns to all inputs. Figure 2 shows the $NHT^O(1)$ set in a 5×5 switch with a speedup of 3. The priority of each input assigned by output 1 is shown next to the input. The arrow points to the $NHT^O(1)$.

Table 1: Output Territory Sets & Input Priorities

	Inputs	Priority(P_j^O)
$NHT^O(j)$	$(j + 1)$ *Mod* N	1
	......	......
	$(j + S)$ *Mod* N	S
$HT^O(j)$	$(j + S + 1)$ *Mod* N	$S + 1$
	......	......
	$(j + N)$ *Mod* N	N

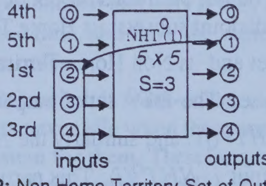


Figure 2: Non Home Territory Set of Output 1 and the Input Priorities Assigned by Output 1 in a 5x5 Switch which has a Speedup of 3.

Definition: *Input Home Territory* (HT^I) *and Non Home Territory* (NHT^I) *sets.*

Each input partitions the set of outputs into two disjoint subsets: its Home Territory (HT^I) set and its Non Home Territory (NHT^I) set. The HT^I set of input i is denoted $HT^I(i)$, and similarly the NHT^I set of input i, $NHT^I(i)$.

Again, the partitioning into Home and Non Home Territory sets is based on priorities. The priority assigned to output j by input i equals the priority assigned to input i by output j. In other words, $P_i^I(j) = P_j^O(i)$. An input then assigns the top S ranked outputs to its NHT^I set and the remaining outputs to its HT^I set.

Table 2 gives the territory sets of input i and lists the priorities it assigns to all outputs. Figure 3 shows the $NHT^I(4)$ set in a 5×5 switch with a speedup of 3. The priority of each output assigned by input 4 is shown next to the output. The arrow points to the $NHT^I(4)$.

Table 2: Input Territory Sets & Output Priorities

	Outputs	Priority(P_i^I)
$NHT^I(i)$	$(i + N - 1)$ *Mod* N	1
	......	......
	$(i + N - S)$ *Mod* N	S
$HT^I(i)$	$(i + N - S - 1)$ *Mod* N	$S + 1$
	......	......
	i *Mod* N	N

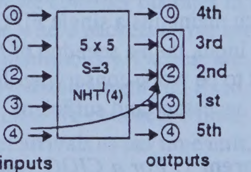


Figure 3: Non Home Territory Set of Input 4 and the Output Priorities Assigned by Input 4 in a 5x5 Switch which has a Speedup of 3.

Definition: *Home Territory VOQs and Non Home Territory VOQs.*

The VOQs at each input are divided into two disjoint subsets according to the NHT^I and HT^I sets of that input, and they are the VOQ_{NHT} set and the VOQ_{HT} set. For an arbitrary input i,

$$VOQ_{NHT}(i) = \bigcup_{j \in NHT^I(i)} \{VOQ_{ij}\}\text{, and}$$

$$VOQ_{HT}(i) = \bigcup_{j \in HT^I(i)} \{VOQ_{ij}\}\text{. Each}$$

VOQ_{ij} has its own unique and fixed priority. We use $P_{VOQ_{ij}}$ to denote the priority of VOQ_{ij}, and $P_{VOQ_{ij}} = P_i^I(j)$.

3.2.2: Home Territory Algorithm

The description of Home Territory Algorithm is divided into three parts. The first part describes the service policy in a single phase. The second part describes how an input selects a granting output. The third part describes the service policy in a single time slot.

Part I. Service Policy in a Phase

The service provided by HTA in a phase has two steps, and they are:

1. conflict free matching between inputs and outputs
2. packet forwarding from inputs to outputs.

HTA finds a conflict free matching in multiple iterations. All inputs and outputs are initially unmatched at the beginning of the first iteration. Only those inputs and outputs that are unmatched at the end of the previous iteration are eligible for matching in subsequent iterations. Iterations are carried out until no further matching is available.

Each iteration has three steps, which operate in parallel on unmatched inputs and unmatched outputs. These three steps are:

1. Each unmatched input sends a request to every output for which it has a queued packet.
2. If an unmatched output receives any requests, it grants to the requesting input with the highest priority.
3. If an unmatched input receives any grants, it accepts an output based on the policy outlined in Part II.

Part II. How an Input Selects a Granting Output

This selection process is carried out in each iteration of every phase. And it has two steps:

1. The input divides its granting outputs into two disjoint sets: the Unserved Set, consisting of outputs that haven't been served by the input in a previous phase of the current time slot; and the Served Set, consisting of outputs that have been served at least once in a previous phase of the current time slot.
2. The input selects a granting output from one of these sets in the following way: If the Unserved Set is non-empty, then the output with the highest priority in that set is selected, otherwise the output with the highest priority in the Served Set is selected.

In other words, an output cannot receive consecutive services at an input, unless all other requesting outputs have been served at least once at this input.

Part III. Service Policy in a Time Slot

HTA carries out the single phase service policy S times per time slot, where S is the speedup of the switch.

3.2.3: Mimicking output-queueing with the HTA

The main result of this section is Theorem 2, which establishes that a CIOQ switch operating under HTA and with speedup bigger than $N/2$ conserves work. We will assume that the switch is completely empty at time $t=0$ and that packets arrive after this time.

Lemma 1: *The VOQ_{NHT}s of all inputs are empty at the end of every time slot. In other words, for any input i, if*

$$\sum_{j \in NHT^I(i)} |VOQ_{ij}(t)| = 0, \text{ then}$$

$$\sum_{j \in NHT^I(i)} |VOQ_{ij}(t+1)| = 0.$$

Proof:

We prove this lemma by induction. Suppose that the VOQ_{NHT} set at each input is empty at the end of time *t*. We will show that if VOQ_{ij} is non-empty at the beginning of time $t+1$, where $P_j^O(i) = k$ and $k \le S$, then it will be empty at the end of time $t+1$.

Suppose VOQ_{ij} isn't served at any phase before Φ_k. Then at Φ_k, input i has the highest priority among all requesting inputs at output j. By the induction hypothesis, all NHT VOQs were empty at the end of time *t*. Given that there is at most one packet arrival per time slot per input, there can be at most 1 packet in each of the NHT VOQs for output *j*. Hence there can be at most $k-1$ competing inputs that have higher priority than input *i* at output j. But, these inputs must have been served during phases Φ_l, $l < k$. Therefore HTA ensures that output j grants input *i*'s request at or before Φ_k. And input i accepts output *j*'s grant, because output j is guaranteed to be in input *i*'s Unserved Set. Hence, if $P_j^O(i) = k$ and $k \le S$, VOQ_{ij} is guaranteed to be cleared/emptied no later than Φ_k. Thus, there will be no packets in VOQ_{NHT}s at the end of time $t+1$. ❐

Lemma 2: *At any time slot, if* $|VOQ_{ij}| \ne 0$ *and* $j \in HT^I(i)$, *then at input i, there can be at most one granting output* j', *such that* $P_i^I(j') > P_i^I(j)$ *and* $j' \in NHT^I(i)$.

Proof:

All VOQ_{NHT}s are empty at the end of every time slot (Lemma1) and at most one packet can arrive at each input per time slot. If the newly arrived packet at input i is destined to output j' and $j' \in NHT^I(i)$, then $P_i^I(j') > P_i^I(j)$. Before input i can accept any granting outputs from its Home Territory set, it must use one phase to clear/empty $VOQ_{ij'}$. Otherwise, $VOQ_{NHT}(i)$ is empty, and input i can start immediately to accept granting outputs from its Home Territory set. ❐

Theorem 2: *A* $N \times N$ *combined input-output queued packet switch using the Home Territory Algorithm is work conserving for all types of input traffic iff its speedup is greater than* $N/2$.

Proof:

Consider any output *j* for which there is a cell at some input. We show that output *j* will be serviced and hence that HTA conserves work. There are essentially two cases to consider.

1. Output *j* has a cell at a NHT VOQ. In this case, Lemma 1 asserts that this cell will be forwarded and j will not idle.

2. The only cell for j is in the HT at some input i. If this is the case, j can loose contention to at most one NHT VOQ at input i (as a consequence of Lemma 2) and to at most $N-S-1$ other HT VOQs at input i. This means it can loose contention in at most $N-S$ phases. But, the speedup $S > N/2$ implies that $S > N-S$. Therefore, there is a phase at which output j will be served at input i, even after all these higher

priority outputs are served.

This proves sufficiency. ❐

To prove the necessity of Theorem 2, we present the following traffic pattern which defeats the Home Territory Algorithm when the speedup is less than or equal to $N/2$.

Given $S = n$, if a $2n \times 2n$ switch fails to conserve work, then a $(2n + 1) \times (2n + 1)$ switch also fails. Therefore, we choose to use a $2n \times 2n$ switch to show the traffic pattern.

The following steps enunciate the proof:

- Consider a $2n \times 2n$ switch operating under HTA.
- Suppose its speedup $S = n$.
- Pick an arbitrary input i.
- Choose VOQ_{iJ} to be the target VOQ, where $P_{VOQ_{iJ}} = 2n$. Observe that we have picked a VOQ which has the lowest priority at input *i*. The traffic pattern has been designed to stress the target VOQ such that at some time one of its packets is required at output *J* and yet can't be forwarded because of HTA, thus causing the switch to fail to conserve work.
- The VOQs at input i are shown in Figure 4. The shaded area represents VOQ_{NHT}, while the non-shaded area represents VOQ_{HT}. The priority of each VOQ is also indicated in the figure.

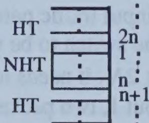


Figure 4: VOQs at input *i*

The following traffic pattern defeats HTA:

1. Identify output *j*, $j \in HT^I(i)$ and $P_{VOQ_{ij}} = n + 1$.
2. Let all inputs receive packets destined to output *j*.
3. Repeat Step 2 until VOQ_{ij} has enough packets to stay non-empty till such time a packet from the target VOQ fails to be delivered to its output. (Note: It is a priori possible to determine how long Step 2 should be repeated to achieve said failure.)
4. Repeat Steps 1 & 2 for all output j's, $j' \in HT^I(i)$ and $n + 2 \le P_{VOQ_{ij'}} \le 2n - 1$.
5. Repeat Step 1 for target VOQ, i.e. VOQ_{iJ}.
6. Now choose an output j'', such that $j'' \in NHT^I(i)$. Let $VOQ_{ij''}$ receive a packet per time slot from now on.
7. Repeat Step 6 until the FIFO queue at output *J* becomes empty.
8. Now output *J* is waiting for the packet at VOQ_{iJ}. However, this packet can't be forwarded according to HTA, since there are n non-empty VOQs at input i that have a higher priorities than the target VOQ. Because $S = n$, the switch fails to conserve work.

This traffic pattern proves necessity. (Note: It may be instructive to construct the traffic pattern taking n=4 and S=2.) ❐

Intuitive Explanation of the Theorem.

For a CIOQ switch that uses HTA to be work conserving, its minimum required speedup is determined by the expression $max(|HT| + 1, |NHT|)$. According to the definition of NHT, $|NHT| = S$. Also during any time slot, to guarantee work conservation, $|HT| + 1$ phases are needed to serve all non-empty VOQs at least once. Hence, it becomes obvious that a speedup of roughly $N/2$ is able to ensure work conservation.

4: Conclusion and Future Work

In this paper, we have proved that a

combined input-output queued switch that uses VOQ buffering can be work-conserving for any type of input traffic if its speedup is greater than $N/2$. This result is proved using a novel scheduling algorithm, the Home Territory Algorithm.

This result only meets part of our original goal, which is to find techniques such that a CIOQ switch can behave identically as an OQ switch. Here, 'behave identically' means their output processes are indistinguishable, given that the input traffic is the same. HTA only guarantees these processes to be busy at the same time. To make the processes indistinguishable, we also need techniques which can guarantee identical packet departure order. This has been our current research focus.

5: References

[1] Karol, M.; Hluchyj, M.; and Morgan, S. "Input versus output queueing on a space division switch," IEEE Trans. Commun., Dec. 1987, vol.COM-35, no.12, p.1347-56.

[2] I. Iliadis and W.E. Denzel, "Performance of packet switches with input and output queueing," in Proc. ICC '90, Atlanta, GA, Apr. 1990. p.747-53.

[3] A.L. Gupta and N.D. Georganas, "Analysis of a packet switch with input and output buffers and speed constraints," in Proc. InfoCom '91, Bal Harbour, FL, Apr. 1991, p.694-700.

[4] Y. Oie; M. Murata, K. Kubota, and H. Miyahara, "Effect of speedup in nonblocking packet switch," in Proc. ICC '89, Boston, MA, Jun. 1989, p. 410-14.

[5] J.S.-C. Chen and T.E. Stern, "Throughput analysis, optimal buffer allocation, and traffic imbalance study of a generic nonblocking packet switch," IEEE J. Select. Areas Commun., Apr. 1991, vol. 9, no. 3, p. 439-49.

Appendix A: Proof of Theorem 1.

A 1: *Speedup Requirement of a Work Conserving Switch that has a Single FIFO Queue at Each Input.*

Theorem 1: *If a $N \times N$ switch maintains a single FIFO queue at each input, then a speedup of N is required for it to be work-conserving.*

To prove this theorem, all we need to show is that such a switch fails to conserve work when $S = N-1$. An input traffic pattern that causes the switch to fail is shown in Figure 5. Assume the scheduler is optimum, i.e. can look infinitely into the future and make the best scheduling decisions possible in an effort to make the switch conserve work.

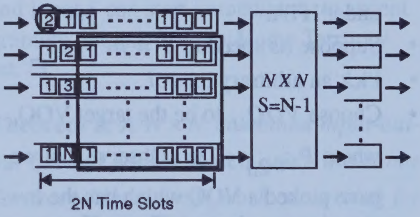


Figure 5: A Counter Example

In Figure 5, each row of packets represents the arrival pattern at an input, with row 1 corresponding to arrivals at input 1, row 2 corresponding to arrivals at input 2, and so on. Each column represents the arrival pattern for all inputs at a single time slot, with the right most column representing arrivals at time slot 1. Packets destined for a particular output are labelled by that output's number. For example, all packets destined to output 1 are labelled 1.

Given the input traffic pattern shown in Figure 5, for the switch to be work conserving at time slot $2N$, it needs to forward $2N$ packets to output 1, two packets to output 2, and one packet to each of the remaining outputs.

In order to conserve work at time $2N-1$, *any* scheduler must serve inputs 2 through N in each preceding time slot. This will allow it to remove all the 1's marked in the rectangle and thus it will conserve work at time $2N-1$. However, this leaves a backlog of packets in input 1 for output 1 (all the 1's in the topmost row), and there will be $2N - 1 - (N-1) = N$ of these 1's at the end of time $2N - 1$.

Therefore, the cell arriving at time $2N$ at input 1, destined for output 2 (shown circled in Figure 5) *cannot* be served, due to HOL blocking.

Hence, *any* schedule is doomed to fail when the speedup is less than N, proving the theorem. ❐

Optimal Space Priority Policies for Shared Memory ATM Systems *

Rajarshi Roy and Shivendra S. Panwar †

Center for Advanced Technology in Telecommunications
Polytechnic University
6 Metrotech Center
Brooklyn, NY 11201

September 20th, 1997

Abstract

In this paper we study the problem of the optimal design of buffer management policies for a shared memory ATM switch or demultiplexer. A system with cells of two different space priorities is considered. Our objective is to determine the optimal policy that minimizes the total weighted cell loss. The problem of finding the optimal policy within the class of *pushout* and *expelling* policies is considered. Using sample path techniques the search space for the optimal policy is reduced to a subset of the entire policy set for each policy class. A numerical study based on the value iteration technique is used to investigate the structure of the optimal policy. This technique is also used to calculate the loss probabilities for different classes of cells for a system with small buffer size. We conclude by proposing heuristic policies suitable for implementation in large buffer systems and also outline future research directions.

Keywords: Markov Decision Theory, Sample path techniques, Shared Memory Switch, Buffer Management, ATM.

1 Introduction

Most of the ATM (Asynchronous Transfer Mode) switch architectures that have been proposed in the literature use some buffering to queue cells whose service has been delayed due to contention for resources within the switch. The location of these buffers and the buffer management policy affects the switch performance i.e., switch throughput and cell delay to a great extent. Performance studies of different types of switches were done previously and it was shown that output queueing with completely shared buffering achieves the optimal throughput-delay performance.[8]

Some buffer management is necessary to ensure that an output-queued shared-memory switch or demultiplexer, both of which have multiple output ports each with its own queue, does not perform poorly when some of the output queues get overloaded. In such a case, without any control the overloaded queue will consume most of the buffer spaces forcing the other lightly loaded queues to incur loss. To ensure efficient sharing of the buffer one needs to develop buffer sharing schemes to ensure that buffer space is shared properly among different logical queues and high priority cells are selectively favoured over low priority cells in case of contention.

---

*This research was supported by the New York State Center for Advanced Technology in Telecommunications, Polytechnic University.

†Tel: 718 260 3740, Fax: 718 260 3074, e-mail: panwar@kanchi.poly.edu

Let us describe the system model and the two different classes of buffer management schemes we will consider. The shared memory switch (Fig. 1) is modelled by a multiserver queue with a bounded buffer. The entire buffer is partitioned into two parts: the main buffer of size B_M and the temporary buffer of size B_T. Time is slotted and transmission of a cell takes one time slot. During one time slot at most B_T cells may arrive to the system and they are placed in the temporary buffer. The cells may belong to different traffic types. This assumption is consistent with the fact that a leaky bucket policed source may inject cells with different priority levels even if they belong to the same virtual circuit. A cost is incurred for each cell that is dropped from the system. The cost is higher for high priority classes. It is an N-ported, slotted system with cells destined to each output port or server constituting a logical queue. FIFO order is to be maintained within each of the logical queues once the cells are accepted to main buffer, each of which contain high and low priority cells. This is because of the fact that cells of the same logical queue may belong to the same virtual circuit. In a switch model reordering of cells is allowed when they are placed from the temporary buffer to the main buffer and this is not allowed in the demultiplexer model. Because in a switch due to same input line speed all the cells which came in the same time slot should belong to different virtual circuits and that may or may not be the case in a demultiplexer model. Service time is deterministic and the servers serve one cell in every time slot if such a cell is available. Clearly, for a switch $N = B_T$, if each input line can inject at most one cell in every time slot. The value of B_T in case of a demultiplexer will be dependent on the input line speed. The objective is to minimize the long run average weighted cost that is incurred from the lost cells. Depending on the available control we have over the dropping of cells from the temporary or main buffer and the placement of cells in the main buffer, we distinguish two classes of policies.

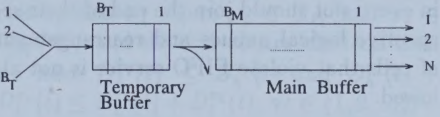


Figure 1: The system model

The first class of policies considered are *pushout* policies. A pushout policy is allowed to discard cells from the main buffer in order to make space available for cells in the temporary buffer which cannot enter the main buffer because it is full. A cell from temporary buffer cannot be blocked from admission to the main buffer if there is space in the main buffer. We narrow down the search space for an optimal pushout policy in a system with two priority classes, with an arbitrary discrete arrival process and with an arbitrary number of logical queues. We prove by sample path techniques that the optimal policy within the pushout class G_P has to lie within the class of *squeeze-out policies* G_{P_1}, which is defined later.

The second class of policies considered are *expelling* policies. Expelling policies are allowed to discard cells from the main buffer or block cells in the temporary buffer from admission into the main buffer irrespective of system state. Properties of the optimal expelling policy are obtained that narrow down the the class of candidate policies considerably in a system with two classes. More specifically speaking, it is shown that according to the optimal *expelling* policy one should apply the *squeeze-out* rule for accommodating cells of each logical queue in the portion of the main buffer allocated to it. While serving the queue, if the cell to be served next is of high priority then it is always served, and if the cell to be served is of low priority then that is either served or all the low priority cells residing at the head of that logical queue are dropped and the first available high priority cell is served.

In both types of policies it is assumed that the cells which enter the main buffer in every slot should join the end of their respective logical queues and rearrangement of cells that violate FIFO service is not allowed.

There has been considerable amount of prior work in this area. The queueing analysis of different buffer sharing schemes and their relative merits was analyzed in [1]. The problem of designing optimal policies for the purpose of optimizing certain performance criteria is considered in [2]. They only consider the problem of searching for optimal policies within the class where packets can only be blocked at the entry point of the buffer; dropping packets once it is accepted in the switch is not allowed. In [3] *pushout* is allowed and arriving packets can push out packets from the longest logical queue. This policy turns out to be optimal only for a system with symmetric arrival and service processes. Guerin, Cidon, Georgiadis and Khamisy [4] considered a Poisson arrival, single loss priority and exponential service model for sharing memory in a switch with two output ports. They have established the optimality of a threshold based pushout scheme using Markov decision theory.They also calculated the values of those thresholds. For a N ported system, $N \geq 3$, they have results only for balanced input rates and equal service rates at all the output ports. They do not consider packets with different loss priorities.

Hung et al in [5] has provided optimal policies within the class of *discarding*, *pushout* and *expelling* policies using dynamic programming and sample path based techniques. They consider the case of general arrival models for the case of a single output port.

Hahne and Choudhury [6] have proposed a set of pushout and buffer sharing rules. They have proposed a backpressure mechanism for sharing buffer across stages and a pushout scheme for sharing buffer among different logical queues sharing a common memory. The lowest priority cell at the head of the queue is selected for pushout. They have applied this work a hierarchical switch system and Banyan based switch system [6, 7].

2 Pushout Policies

The class of *pushout* policies G_P is defined by the following rules: (a) A cell can be expelled from the main buffer only if it is pushed out by another cell in the temporary buffer and the main buffer is full, (b) A cell from the temporary buffer can be discarded only if the main buffer is full, (c) A cell can never be dropped if there is room for it in the main buffer.

The class of policies G_{P_1} is defined as the following: Suppose we are given the number of buffer positions that will be allocated to a particular logical queue after the drop/pushout decision at a particular integer time in the main buffer is strictly less than the number of cells it had in the entire system just before the decision epoch t, i.e. at t^- a given number of cells are to be dropped/pushed out from each logical queue. Then the rules to be followed are (a) Append the cells that are in the temporary buffer and belong to logical queue n to the end of the main buffer in the allocated position, high priority cells first (if switch model) or in FIFO order (if demultiplexer model). (b) If the amount of buffer allocated for logical queue n in the main buffer is full, and there are cells of that logical queue in the temporary buffer, push out the low-priority cells starting from those closest to the head of that logical queue, (c) If all low priority cells of a logical queue are pushed out, discard all the remaining cells of that logical queue which are in the temporary buffer. This entire procedure defines the *squeeze-out* class of policies.

2.1 Problem formulation

Cells are classified into two priority classes. The high priority class is more sensitive to cell losses. Without loss of generality we assume that the priority of class 1 is higher than the priority of class 2. The priority of a class is reflected by the cost C_h and C_l that is incurred by the dropping of a cell of high and low priority respectively. We denote by $X_{n,i}(t)$ the class of the cell residing at the i-th place of the n-th logical queue by the end of the slot $[t-1,t]$. Here $i = 1,2,...,L_n(t)$, where $L_n(t)$ is the number of buffer places the logical queue n occupies at t^-. Clearly, $\sum_{n=1}^{N} L_n(t) = B$. $X_1(t), X_2(t), ..., X_N(t)$ is the vector that can be taken as a state, where $X_n(t) = (X_{n,i}(t) : i = 1,2,....,L_n(t))$. If there are any empty buffer places, they can be arbitrarily assumed to be part of any particular fixed logical queue and the value of $X_{n,i}(t)$ corresponding to them is 0. We want to find out the policy π_1 amongst the class of *pushout* policies G_P so that $E_{\pi_1}(\sum_{t=0}^{\infty}(C_h \cdot D_h(t) + C_l \cdot D_l(t))|\text{any initial state } = x) \leq E_{\pi_2}(\sum_{t=0}^{\infty}(C_h \cdot D_h(t) + C_l \cdot D_l(t))|\text{any in initial state} = x)$, $\forall \pi_2 \in G_P$. Here, $D_h(t)$ and $D_l(t)$ are the number of high and low priority cells dropped from the system up to the end of slot $[t-1,t]$, respectively. C_h and C_l are positive real numbers, $C_h \geq C_l$. E_{π_i} the expectation when policy π_i is used. One can write with this value function the dynamic programming equations if the arrival statistics are completely characterized [10].

The objective of Theorem 1, which is stated below, is to show that the policy π^{po} which is optimal in G_P in the sense that $D_h^{\pi^{po}}(t)+D_l^{\pi^{po}}(t) \leq D_h^{\pi}(t)+D_l^{\pi}(t)$, $\forall$ integer time $t \in \{0,1,2,3,....\}$ and $\forall \pi \in G_P$ such that $\pi \neq \pi_{po}$ and also $D_h^{\pi^{po}}(t) \leq D_h^{\pi}(t)$; lies in $G_{P_1} \subset G_P$. Here, $D_h^{\pi}(t)$ is the number of high priority cells dropped up to time t under policy π and $D_l^{\pi}(t)$ is the number of low priority cells dropped up to time t under that policy.

Theorem 1: When the systems start with the same initial conditions and under policies π_1 and π_2 where $\pi_1 \in G_{P_1}$ and $\pi_2 \in G_{P_1}^c \cap G_P$ and the arrivals are identical then we have $D_h^{\pi_1}(t) \leq D_h^{\pi_2}(t)$ and $D_h^{\pi_1}(t) + D_l^{\pi_1}(t) \leq D_h^{\pi_2}(t) + D_l^{\pi_2}(t)$, $\forall t \in \{1,2,...\}$.

The proof of this theorem depends on Lemma 1.1 which is stated below.

Lemma 1.1: For any policy $\pi_2 \in G_{P_1}^c \cap G_P$ there exists a policy π_3, which acts similarly as a policy that belongs to G_{P_1} at $t = 1$, and if appropriately defined for all integer time $t \geq 1$, under the same initial condition and set of arrivals $D_h^{\pi_3}(t) \leq D_h^{\pi_2}(t)$ and $D_h^{\pi_3}(t) + D_l^{\pi_3}(t) \leq D_h^{\pi_2}(t) + D_l^{\pi_2}(t)$, $\forall t \in \{1,2,...\}$.

The proof of Lemma 1.1 and Theorem 1 is not given here due to space restrictions. [9]

So for any policy π', there exists a policy $\pi_1 \in G_{P_1}$ such that $D_l^{\pi_1}(t) \geq D_l^{\pi'}(t)$ and

$$D_h^{\pi_1}(t) + D_l^{\pi_1}(t) = D_h^{\pi'}(t) + D_l^{\pi'}(t)$$

$$D_l^{\pi_1}(t) - D_l^{\pi'}(t) = D_h^{\pi'}(t) - D_h^{\pi_1}(t) \geq 0$$

$$\begin{aligned} 0 &\leq C_l(D_l^{\pi_1}(t) - D_l^{\pi'}(t)) \\ &\leq C_h(D_h^{\pi'}(t) - D_h^{\pi_1}(t)) \end{aligned}$$

$$\begin{aligned} 0 &\leq C_l \cdot D_l^{\pi_1}(t) + C_h \cdot D_h^{\pi_1}(t) \\ &\leq C_l \cdot D_l^{\pi'}(t) + C_h \cdot D_h^{\pi'}(t) \end{aligned}$$

So,

$$E_{\pi_1}[\sum_{t=0}^{N}(C_h \cdot D_h(t) + C_l \cdot D_l(t))|$$

$$\text{any initial state} = x)] \leq$$

$$E_{\pi'}[\sum_{t=0}^{N}(C_h \cdot D_h(t) + C_l \cdot D_l(t))|$$

$$\text{any initial state} = x]$$

Thus Theorem 1 is equivalent to proving the optimality of the class of policies G_{P_1}.

2.2 Some Counterexamples

These are counterexamples to two sample-path dominance conjectures. These conjectures are true in expected value sense at least for certain values of the load parameters. That fact is revealed by the numerical results presented in the next section.

2.2.1 Counterexample 1

This is a counterexample to the conjecture that a high priority cell should always pushout a low priority cell, if there is any. Here it will be shown that a performance criteria such as given the previous section cannot be satisfied for every sample path of the input process if such a policy is followed. We consider a 2x2 switch with four main buffer places and two temporary buffer places. We refer to the policy that recommends pushout from the longest queue with low priority cell as the conjectured policy and the other one as an alternative policy. We start with the same initial state with one low priority cell destined to output 1, three high priority cells destined to output 2 and two high priority cells in the temporary buffer destined to output 2. Under the conjectured policy the low priority cell is dropped and one of the high priority cells are accepted from the temporary buffer. Under the alternative policy both the high priority cells are dropped from the temporary buffer. Now if two high priority cells come in the next time slot then at the end of the second slot under both the policies we have the same state. If in all the future decision epochs we follow same actions for both the policies, then under the alternative policy we incur less total packet loss.

2.2.2 Counterexample 2

This is a counterexample of the conjecture that a low priority cell can never push out a high priority cell under the optimal policy and this is true for every sample path. Consider a system with main buffer size of four and temporary buffer size of two. The initial state is such that there are four high priority cells destined to output 2 in the main buffer and one low priority cell destined to output 1 in the temporary buffer. Under the conjectured policy we drop the low priority cell and under the alternative policy we pushout one high priority cell from the main buffer and admit the low priority cell. Under the conjectured policy we serve one high priority cell destined to output 2 and under the alternative policy we serve the low priority cell destined to output 1 and one high priority cell destined to output 2. If in the next slot two high priority cells destined to output 2 arrive, then at the end of that slot we enter the same state under both the policies. In future decision epochs both the policies can take the same actions at every slot. So the total cell loss under the alternative policy will be less.

These counterexamples show that inter queue pushout policy investigation is more complex and the results are not sample path wise true but may be true in the expected value sense. We therefore did a numerical study using value iteration. The goal of the ongoing work is to find out analytical structures of the optimal policy using the numerical study as a supporting tool.

2.3 Numerical study for Pushout Policy

We considered a two-ported shared memory switch modelled as a queue with bounded buffer and two servers each with a constant service rate of one cell/time slot. At any time slot at most two cells can come into the system. The arrival process is i.i.d. and the cells can belong to either of the two logical queues with probabilities b_1 and b_2. Given the event that a cell belongs to logical queue 1, it can be of high or low priority with probability c_1 and c_2 and given the event that it belongs to logical queue 2, it can be of high or low priority with probability d_1 and d_2. We have here $b_1 + b_2 = 1$, $c_1 + c_2 = 1$ and $d_1 + d_2 = 1$. Our system has six buffer places of which two are temporary buffer places. At every decision epoch we have to make sure that after the cell dropping/pushout decision is taken we should not be left with more than four cells in the buffer.

For our numerical study we have assumed $B_M = 4$ and the arrival process is i.i.d. with the statistics that at any time slot 0, 1 or 2 cells can come into the system with probabilities a_0, a_1 and a_2 respectively. The sum of these probabilities is 1. We formulate the problem in the framework of Markov decision theory and seek to find out the optimal policy numerically that minimizes the total undiscounted expected cost.

The numerical results reveal that pushout from the longest queue policy when we have only either high or low priority cells in the buffer is not optimal for an unbalanced load. Rather, the logical queues have thresholds k_1 and k_2 with their sum at least B_M. Thus if we need to drop cells we should check which logical queue has exceeded its threshold. If the new cells belong to that queue then they get dropped. If it is the other one then these new cells are accepted and cells are pushed out from the other logical queue to make place for them. For a balanced load when there are a total of three cells of high priority for both the logical queues, numerical computation shows that we should drop one from each logical queue. But for the case of 70 percent of the incoming traffic going to output port 1, we see that we should drop 2 cells from logical queue 1. This clearly shows the change of value of thresholds with the offered load.

3 Expelling Policies

The class of *expelling* policies G^E has as members all policies that append the new cells from the temporary buffer and do not rearrange them. This class of policies is even allowed to drop cells from the main buffer when it is not full. The only constraint is that FIFO service order must be maintained.

In the following we will prove that the optimal policy within the class G^E belongs to a subset of that class G^{EO}.

The class G^{EO} is defined as the following:

1. Cells from each logical queue are moved from the temporary buffer to the portion of the main buffer allocated for it in that slot, high priority cells first. If they do not fit then low priority cells are expelled starting from those closest to the head of the queue.

2a. If the cell at the head of the queue is of high priority then it is served.

2b. If the cell at the head of the queue is of low priority then either that cell is served or all the low priority cells from the head of the queue until the high priority cell closest to the head of the queue are expelled and that high priority cell is served.

Theorem 2: For every policy $\pi \in G^E \cap G^{EO^c}$ there exists a policy $\pi_1 \in G^{EO}$ such that if the system starts from the same initial state under the two policies and the arrival process is identical under the two poli-

cies we have
$D^{1h}(t) \leq D^h(t)$
$D^{1h}(t) + D^{1l}(t) \leq D^h(t) + D^l(t), t = 1, 2, ...$
where $D^{1h}(t)$,$D^{1l}(t)$ is the number of lost cells of high and low priority respectively under π_1 and similarly for $D^h(t)$, $D^l(t)$ under π.

The proof of this theorem requires the following lemma.

Lemma 2.1: For every policy $\pi \in G^E$ there exists a policy π_2, which acts according to rules 1 and 2 and is defined appropriately for $t = 2, 3, ...,$, based on π, so that if the system starts from the same initial state and the arrival process is same under the two policies, then we have

$D^{2h}(t) \leq D^h(t)$
$D^{2h}(t) + D^{2l}(t) \leq D^h(t) + D^l(t), t = 1, 2, ...$
where $D^{2h}(t)$,$D^{2l}(t)$ is the number of lost cells of high and low priority respectively under π_2 and similarly for $D^h(t)$, $D^l(t)$ under π.

The proof of the Lemma 2.1 and Theorem 2 is not given here due to space restrictions [9].

3.1 Numerical study for Expelling Policy

We used the same system as that described in the numerical study for pushout policies. We confirmed that at certain states all the low priority cells at the head of a logical queue are dropped and the high priority cell is served. The decision of whether to serve the low priority cell or to drop them depends on the number of high priority cells of that logical queue in the system for most of the cases. However, we did see that there are some pair of states such that both have more than one low priority cell at the head of the queue and with same number of high and low priority cells in the system, but the action chosen is different.

Blocking probability calculations using value iteration shows that for the same loss probability expelling policies can handle a higher load region than pushout policies. Expelling policy does allow us to improve the performance of high priority class compared to pushout policies at the cost of some degradation in the performance of low priority class. Here the gain of the Markov decision process is calculated, which gives us the loss probabilities of each class of cells [11]. In the results presented in the tables $P(H_i)$ and $P(L_i)$ indicate the loss probability of high and low priority cells of the i-th logical queue and $P(L_i)$ denotes the loss probability of the low priority cells of the i-th logical queue, respectively. C_h is varied to achieve different degrees of expelling low priority cells in favor of high priority cells. C_l, the cost of dropping a low priority cell is always 1.

4 Conclusion

In this paper we partially characterize the structure of the optimum *pushout* and *expelling* policies. Our future goal is to develop heuristics for the problem. We intend to find out heuristic policies for sharing the buffer among different logical queues and apply *squeeze-out* and a threshold based *expelling* policy for every logical queue. The performance for this policy for shared memory switches with large buffer size and bursty arrival process input will be studied using simulations. A simple policy can be used for buffer sharing among different logical queues. One can think of pushout from the longest queue with low priority cells policy as one such candidate policy. Under this policy if no low priority cell is present in the system we can pick the longest queue for pushout. If we think of some kind of pushout with threshold policy then pushout from the queue with largest difference between its present queue length and its threshold can be considered. In all such cases one can additionally apply *expelling* policies depending on some threshold on the number of high priority cells in that logical queue. The numerical results for small buffer sizes and i.i.d. arrival process input indicates that the *expelling* policy can improve the overall performance and the high priority loss performance. Similar results are expected from the proposed simulation study for large buffer size.

References

[1] F. Kamoun and L. Kleinrock, "Analysis of Shared Finite Storage in a Computer Network Node Environment Under General Traffic Conditions," *IEEE Trans. on Communications*, Vol. COM-28, No. 7, July 1980.

[2] G. J. Foschini and B. Gopinath, "Sharing Memory Optimally," *IEEE Trans. on Communications*, Vol. COM-31, No. 3, March 1983.

[3] S. X. Wei, E. J. Coyle and M. T. Hsiao, "An Optimal Buffer Management Policy for High Performance Packet Switching," *IEEE GLOBECOM'91*, Vol. 2, pp. 924-928, December, 1991.

[4] I. Cidon, L.Georgiadis, R. Guerin, A. Khamisy, "Optimal buffer sharing," *IEEE JSAC*, September 1995, vol 13, No. 7, pp. 1229-1240.

[5] L. Tassiulas, Y. C. Hung and S. S. Panwar, "Optimal Buffer Control During Congestion in an ATM Network Node," *IEEE/ACM Transactions on Networking*, August 1994, Vol. 2, No. 4, pp. 374-386.

[6] Abhijit Choudhury, Ellen L. Hahne, "Buffer Management in a Hierarchical Shared Memory Switch," *IEEE INFOCOM'94*, Vol. 3, pp. 1410-1419, June, 1994.

[7] Debasis Basak, Abhijit K. Choudhury, Ellen L. Hahne, "Sharing Memory in Banyan-based ATM Switches," *IEEE JSAC*, vol. 15, No. 5, pp. 881-891, June 1997.

[8] Mark J. Karol, Michael J. Hluchyj and Samuel P. Morgan, "Input Versus Output Queueing on a Space-Division Packet Switch," *IEEE Transactions on Communications*, December 1987, Vol. 35, No. 12, pp. 1347-1356.

[9] Rajarshi Roy and Shivendra S. Panwar, "Optimal Space Priority Policies for Shared Memory ATM Switch or Demultiplexer," *Submitted to INFOCOM'98.*

[10] S. M. Ross, *"Introduction to Stochastic Dynamic Programming"* , Academic Press, New York, 1983.

[11] A. R. Odoni, "On finding the maximal gain for Markov decision processes," *Operations Research*, 17(1969), 857-860.

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^1$	$C_h = 10^2$	$C_h = 10^3$
$P(H_1)$	$3.90\text{x}10^{-4}$	$3.85\text{x}10^{-4}$	$3.24\text{x}10^{-4}$	$2.20\text{x}10^{-4}$
$P(L_1)$	$3.0\text{x}10^{-3}$	$4.0\text{x}10^{-4}$	$1.79\text{x}10^{-3}$	$1.01\text{x}10^{-2}$
$P(H_2)$	$3.90\text{x}10^{-4}$	$3.85\text{x}10^{-4}$	$3.24\text{x}10^{-4}$	$2.20\text{x}10^{-4}$
$P(L_2)$	$3.0\text{x}10^{-3}$	$4.0\text{x}10^{-4}$	$1.79\text{x}10^{-3}$	$1.01\text{x}10^{-2}$

Table 1: Offered Load=0.5; Splitting of Traffic=1:1; High:Low=9:1

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$1.02\text{x}10^{-5}$	$4.0\text{x}10^{-6}$	$1.13\text{x}10^{-6}$	$1.13\text{x}10^{-6}$
$P(L_1)$	$5.10\text{x}10^{-4}$	$5.0\text{x}10^{-2}$	$3.09\text{x}10^{-2}$	$3.09\text{x}10^{-2}$
$P(H_2)$	$1.02\text{x}10^{-5}$	$4.0\text{x}10^{-6}$	$1.13\text{x}10^{-6}$	$1.13\text{x}10^{-6}$
$P(L_2)$	$5.10\text{x}10^{-4}$	$5.0\text{x}10^{-2}$	$3.09\text{x}10^{-2}$	$3.09\text{x}10^{-2}$

Table 2: Offered Load=0.5; Splitting of Traffic=1:1; High:Low=1:1

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$3.0\text{x}10^{-9}$	$6.0\text{x}10^{-10}$	$3.10\text{x}10^{-11}$	$2.0\text{x}10^{-12}$
$P(L_1)$	$6.0\text{x}10^{-4}$	$6.10\text{x}10^{-4}$	$7.02\text{x}10^{-4}$	$1.0\text{x}10^{-2}$
$P(H_2)$	$3.0\text{x}10^{-9}$	$6.0\text{x}10^{-10}$	$3.10\text{x}10^{-11}$	$2.0\text{x}10^{-12}$
$P(L_2)$	$6.0\text{x}10^{-4}$	$6.10\text{x}10^{-4}$	$7.02\text{x}10^{-4}$	$1.0\text{x}10^{-2}$

Table 3: Offered Load=0.5; Splitting of Traffic=1:1; High:Low=1:9

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$5.0\text{x}10^{-4}$	$1.01\text{x}10^{-4}$	$4.0\text{x}10^{-5}$	$4.0\text{x}10^{-5}$
$P(L_1)$	$1.7\text{x}10^{-2}$	$4.7\text{x}10^{-2}$	$1.0\text{x}10^{-1}$	$1.0\text{x}10^{-1}$
$P(H_2)$	$3.0\text{x}10^{-7}$	$7.0\text{x}10^{-9}$	$1.0\text{x}10^{-9}$	$1.0\text{x}10^{-9}$
$P(L_2)$	$2.4\text{x}10^{-4}$	$4.1\text{x}10^{-5}$	$4.0\text{x}10^{-3}$	$4.0\text{x}10^{-3}$

Table 4: Offered Load=0.5; Splitting of Traffic=4:1; High:Low=1:1

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$1.0\text{x}10^{-7}$	$2.0\text{x}10^{-8}$	$1.1\text{x}10^{-8}$	$4.0\text{x}10^{-11}$
$P(L_1)$	$1.8\text{x}10^{-2}$	$1.9\text{x}10^{-2}$	$2.0\text{x}10^{-2}$	$6.0\text{x}10^{-2}$
$P(H_2)$	$7.0\text{x}10^{-11}$	$4.0\text{x}10^{-12}$	$1.0\text{x}10^{-13}$	$8.01\text{x}10^{-15}$
$P(L_2)$	$7.6\text{x}10^{-6}$	$6.0\text{x}10^{-7}$	$4.0\text{x}10^{-6}$	$1.5\text{x}10^{-3}$

Table 5: Offered Load=0.5; Splitting of Traffic=4:1; High:Low=1:9

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^1$	$C_h = 10^2$	$C_h = 10^3$
$P(H_1)$	$5.3\text{x}10^{-3}$	$4.32\text{x}10^{-3}$	$3.65\text{x}10^{-3}$	$3.60\text{x}10^{-3}$
$P(L_1)$	$3.51\text{x}10^{-3}$	$9.02\text{x}10^{-3}$	$2.87\text{x}10^{-2}$	$2.82\text{x}10^{-2}$
$P(H_2)$	$5.3\text{x}10^{-3}$	$4.32\text{x}10^{-3}$	$3.65\text{x}10^{-3}$	$3.60\text{x}10^{-3}$
$P(L_2)$	$3.51\text{x}10^{-3}$	$9.02\text{x}10^{-3}$	$2.87\text{x}10^{-2}$	$2.82\text{x}10^{-2}$

Table 6: Offered Load=0.7; Splitting of Traffic=1:1; High:Low=9:1

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$2.54\text{x}10^{-4}$	$6.32\text{x}10^{-5}$	$2.0\text{x}10^{-5}$	$2.0\text{x}10^{-5}$
$P(L_1)$	$8.45\text{x}10^{-3}$	$1.60\text{x}10^{-2}$	$7.58\text{x}10^{-2}$	$7.58\text{x}10^{-2}$
$P(H_2)$	$2.54\text{x}10^{-4}$	$6.32\text{x}10^{-5}$	$2\text{x}10^{-5}$	$2\text{x}10^{-5}$
$P(L_2)$	$8.45\text{x}10^{-3}$	$1.60\text{x}10^{-2}$	$7.58\text{x}10^{-2}$	$7.58\text{x}10^{-2}$

Table 7: Offered Load=0.7; Splitting of Traffic=1:1; High:Low=1:1

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$4.75\text{x}10^{-8}$	$1.11\text{x}10^{-8}$	$6.23\text{x}10^{-10}$	$2.49\text{x}10^{-11}$
$P(L_1)$	$8.58\text{x}10^{-3}$	$8.81\text{x}10^{-3}$	$9.40\text{x}10^{-3}$	$3.82\text{x}10^{-2}$
$P(H_2)$	$4.75\text{x}10^{-8}$	$1.11\text{x}10^{-8}$	$6.23\text{x}10^{-10}$	$2.49\text{x}10^{-11}$
$P(L_2)$	$8.58\text{x}10^{-3}$	$8.81\text{x}10^{-3}$	$9.40\text{x}10^{-3}$	$3.82\text{x}10^{-2}$

Table 8: Offered Load=0.7; Splitting of Traffic=1:1; High:Low=1:9

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$6.40\text{x}10^{-3}$	$6.70\text{x}10^{-4}$	$6.70\text{x}10^{-4}$	$6.70\text{x}10^{-4}$
$P(L_1)$	$1.56\text{x}10^{-1}$	$2.68\text{x}10^{-1}$	$2.68\text{x}10^{-1}$	$2.68\text{x}10^{-1}$
$P(H_2)$	$1.82\text{x}10^{-7}$	$1.68\text{x}10^{-8}$	$1.68\text{x}10^{-8}$	$1.68\text{x}10^{-8}$
$P(L_2)$	$2.81\text{x}10^{-3}$	$8.24\text{x}10^{-3}$	$8.24\text{x}10^{-3}$	$8.24\text{x}10^{-3}$

Table 9: Offered Load=0.7; Splitting of Traffic=4:1; High:Low=1:1

Loss Probabilities	Pushout Policy	Expelling Policy		
		$C_h = 10^3$	$C_h = 10^6$	$C_h = 10^{11}$
$P(H_1)$	$9.63\text{x}10^{-7}$	$2.85\text{x}10^{-7}$	$1.44\text{x}10^{-7}$	$4.75\text{x}10^{-10}$
$P(L_1)$	$1.63\text{x}10^{-1}$	$1.64\text{x}10^{-1}$	$1.68\text{x}10^{-1}$	$2.15\text{x}10^{-1}$
$P(H_2)$	$5.76\text{x}10^{-11}$	$5.73\text{x}10^{-12}$	$1.20\text{x}10^{-12}$	$9.0\text{x}10^{-14}$
$P(L_2)$	$3.01\text{x}10^{-6}$	$7.64\text{x}10^{-6}$	$1.53\text{x}10^{-5}$	$2.96\text{x}10^{-3}$

Table 10: Offered Load=0.7; Splitting of Traffic=4:1; High:Low=1:9

Necessary and Sufficient Conditions for a Semi-interval Polytope, with Odd Part Fixed, to Contain a Hurwitz Polynomial

L. R. Pujara
Department of Electrical Engineering
Wright State University
Dayton OH 45435
Email: rpujara@cs.wright.edu

Abstract

In this paper, a weaker version of the Interval Polytope Problem is completely solved. The Interval Polytope Problem is stated as follows: given a polytope of polynomials generated by an interval polynomial, what are the necessary and sufficient conditions for it to contain a Hurwitz polynomial? In this paper, a weaker version of the problem is completely solved; i.e., a set of necessary and sufficient conditions are established for an interval polytope with its odd part fixed, to contain a Hurwitz polynomial. The set of conditions turn out to be extremely simple to check. In fact, the question of the existence of the solution of a Hurwitz polynomial is shown to be equivalent to the existence of a special type of solution of a linear programming problem.

Introduction

The general question of establishing necessary and sufficient conditions for an interval polytope of polynomials to contain a Hurwitz polynomial is pretty hard and is still open. A polytope generated by an interval polynomial is called an interval polytope. The deep interest in this problem is motivated by several important control synthesis problems. For example, given a single-input single-output plant, what are the necessary and sufficient conditions for the existence of a fixed order controller that will stabilize the plant? Another example is to determine necessary and sufficient conditions for the existence of a constant output feedback for any SIMO or MISO plant to stabilize the plant. A few simple calculations would immediately show that these two problems are equivalent to determining necessary and sufficient conditions for a polytope of polynomials to contain a Hurwitz polynomial. We will call it the Polytope Problem. Thus, the Interval Polytope Problem is a restricted version of the Polytope Problem where the polytope is an interval polytope. Thus, the present paper is motivated by such synthesis considerations.

In [3], Prasad and Hollot proved that the Hurwitz region is unbounded, connected and in an interval polytope, it intersects at least one of the proper faces of maximum dimension. In [5], the author proved that an interval polytope of degree four has a Hurwitz polynomial if and only if the Kharitonov rectangle of a suitable subpolytope contains a Hurwitz polynomial. In the same paper, the author also noted that if an interval polytope of degree 3 has a Hurwitz polynomial, then one of the four Kharitonov polynomials is Hurwitz. Polytope related issues have been extensively covered in two recent books [[1], [2]] by eminent researchers. The results of the author in [7] are also inspired by questions related to this paper.

In this paper, a set of necessary and sufficient conditions for a semi-interval polytope, (whose odd part is fixed), to contain a Hurwitz polynomial are established. In the view of the author, the approach adopted in this paper seems to have potential to solve the general interval polytope problem.

In this paper, for the sake of convenience of notation, the theory will be explained for a ninth degree interval polynomial. From the exposition, it will be clear that the arguments are valid for any interval polynomial of any degree.

To illustrate the results, a numerical example will be presented.

Problem Formulation

First, a few definitions and notations are presented.

Definition 1. A polynomial of the type

$$a_n + a_{n-1}s + a_{n-2}s^2 + \cdots + a_0 s^n \tag{1}$$

where

$$a_i^- \le a_i \le a_i^+, a_i^- > 0, \quad 0 \le i \le n \tag{2}$$

is called an interval polynomial and the polytope $\mathcal{P}$ generated by it is called an interval polytope.

Definition 2. A polytope $\mathcal{P}_e$ generated by (1) is called semi-interval if its odd part is fixed and the even part is variable.

The problem, which is the main focus of this paper, can be stated as follows: what are the necessary and sufficient conditions for $\mathcal{P}_e$ to contain a Hurwitz polynomial?

The Main Results

Before we prove the main result, some preliminary results are needed which lay the foundation for the main results.

As was mentioned earlier, we will give the exposition of the results for n=9 and from the nature of the exposition, it will be clear that the results are true for any n.

For n=9, the general semi-interval polynomial with odd part fixed can be written as

$$a_9 + b_8 s + a_7 s^2 + b_6 s^3 + \cdots + a_1 s^8 + b_0 s^9 \tag{3}$$

$$0 < a_i^- \le a_i \le a_i^+, \ i = 1,3,5,7,9.$$

Write

$$f^o(p) = b_8 - b_6 p + b_4 p^2 - b_2 p^3 + b_0 p^4 \tag{4}$$

where $p = -s^2$.

Note that $f^o(p)$ is related to the odd part of (3) in the obvious way. We will denote by $\mathcal{P}_e$ the semi-interval polytope generated by (3).

Write
$$m_e(p) = a_9^- - a_7^+ p + a_5^- p^2 - a_3^+ p^3 + a_1^- p^4 \tag{5}$$

and

$$M_e(p) = a_9^+ - a_7^- p + a_5^+ p^2 - a_3^- p^3 + a_1^+ p^4 \tag{6}$$

where $p = -s^2$.

It is clear that for the semi-interval polytope $\mathcal{P}_e$ (generated by (3) with odd part fixed) to contain a Hurwitz polynomial, it is necessary that all the roots of (4) be positive and distinct. The four roots of (4) are denoted by O_1, O_2, O_3, O_4 ordered as

$$0 < O_1 < O_2 < O_3 < O_4. \tag{6'}$$

Define

$$A_i = m_e(O_i) \text{ and } B_i = M_e(O_i), \quad 1 \le i \le 4. \tag{7}$$

It is obvious that a necessary condition for $\mathcal{P}_e$ to contain a Hurwitz polynomial is that

$$A_1 < 0, \ A_3 < 0; \ B_2 > 0, \ B_4 > 0. \tag{8}$$

Throughout this paper, it will be assumed that the conditions in (8) hold and the roots of the odd part satisfy (6′).

Definition 3.

$$H_{\ell m} = A_\ell B_m - A_m B_\ell, \quad \ell = 1,3 \quad m = 2,4. \tag{9}$$

Theorems 1 and 2 are quite basic for this paper.

Theorem 1. For the semi-interval polytope $\mathcal{P}_e$, $H_{\ell m}$ is given by

$$\begin{aligned} H_{\ell m} &= (A_\ell - A_m)\Delta_9 + (O_m A_\ell - O_\ell A_m)\Delta_7 + (O_m^2 A_\ell - O_\ell^2 A_m)\Delta_5 \\ &\quad + (O_m^3 A_\ell - O_\ell^3 A_m)\Delta_3 + (O_m^4 A_\ell - O_\ell^4 A_m)\Delta_1 \end{aligned} \tag{10}$$

where

$$\Delta_i = a_i^+ - a_i^-, \quad i = 1,3,5,7,9. \tag{11}$$

Also,

$$\begin{aligned} H_{\ell m} &= (B_\ell - B_m)\Delta_9 + (O_m B_\ell - O_\ell B_m)\Delta_7 + (O_m^2 B_\ell - O_\ell^2 B_m)\Delta_5 \\ &\quad + (O_m^3 B_\ell - O_\ell^3 B_m)\Delta_3 + (O_m^4 B_\ell - O_\ell^4 B_m)\Delta_1 \end{aligned} \tag{12}$$

where $\ell = 1,3;\ m = 2,4.$

(10) will be called the lower decomposition and (12) will be called the upper decomposition of $H_{\ell m}$.

Proof. First observe that

$$B_r - A_r = \Delta_9 + O_r\Delta_7 + O_r^2\Delta_5 + O_r^3\Delta_3 + O_r^4\Delta_1, r = 1,2,3,4.$$

Now

$$\begin{aligned} H_{\ell m} &= A_\ell B_m - A_m B\ell \\ &= A_\ell(B_m - A_m) - A_m(B_\ell - A_\ell) \\ &= A_\ell[\Delta_9 + O_m\Delta_7 + O_m^2\Delta_5 + O_m^3\Delta_3 + O_m^4\Delta_1] \\ &\qquad - A_m[\Delta_9 + O_\ell\Delta_7 + O_\ell^2\Delta_5 + O_\ell^3\Delta_3 + O_\ell^4\Delta_\ell] \\ &\quad = (A_\ell - A_m)\Delta_9 + (O_m A_\ell - O_\ell A_m)\Delta_7 \\ &\quad + (O_m^2 A_\ell - O_\ell^2 A_m)\Delta_5 + (O_m^3 A_\ell - O_\ell^3 A_m)\Delta_3 + (O_m^4 A_\ell - O_\ell^4 A_m)\Delta_1 \end{aligned}$$

This proves (10). The proof of (12) is similar.

QED

Note 1. As usual [1], the convex hull of the four Kharitonov polynomials of an interval polytope is called the Kharitonov rectangle. Note that for the semi-interval polytope $\mathcal{P}_e$, the odd part being fixed, the Kharitonov rectangle degenerates into a straight line and it will be called the Kharitonov line.

Next, a set of necessary and sufficient conditions are presented which ensure the existence of a Hurwitz polynomial in the Kharitonov line of $\mathcal{P}_e$.

Theorem 2. The Kharitonov line of the semi-interval polytope contains a Hurwitz polynomial if and only if

$$\text{(a)}\ A_1 < 0,\ \ A_3 < 0;\ \ B_2 > 0,\ B_4 > 0 \tag{13}$$

$$\text{(b)}\ H_{\ell m} < 0;\ \ \ell = 1,3;\ \ m = 2,4. \tag{14}$$

Proof. Suppose the Kharitonov line of $\mathcal{P}_e$ contains a Hurwitz polynomial. Call it g(s). Therefore, there exists λ, $0 \leq \lambda \leq 1$ such that the even part $g^e(s)$ of g(s) is given by

$$g^e(p) = \lambda M_e(p) + (1-\lambda)m_e(p) \quad \text{where} \quad p = -s^2. \tag{15}$$

Since g(s) is stable, it follows that

$$g^e(O_1) < 0.\ \text{This implies that}\ \lambda B_1 + (1-\lambda)A_1 < 0 \Rightarrow \lambda < \frac{-A_1}{B_1 - A_1}. \tag{16}$$

Similarly,

$$g^e(O_2) > 0 \Rightarrow \lambda > \frac{-A_2}{B_2 - A_2}. \tag{17}$$

$$g^e(O_3) < 0 \Rightarrow \lambda < \frac{-A_3}{B_3 - A_3}. \tag{18}$$

$$g^e(O_4) > 0 \Rightarrow \lambda > \frac{-A_4}{B_4 - A_4}. \tag{19}$$

From (16) and (17), it follows that

$$\frac{-A_2}{B_2 - A_2} < \frac{-A_1}{B_1 - A_1} \tag{20}$$

which implies that

$$H_{12} = A_1B_2 - A_2B_1 < 0.$$

Other conditions on $H_{\ell m}$ are established similarly.

Conversely, assume that the conditions of (14) hold. Now,

$$H_{12} = A_1B_2 - A_2B_1 < 0 \Rightarrow A_1(B_2 - A_2) - A_2(B_1 - A_1) < 0 \Rightarrow \frac{-A_2}{B_2 - A_2} < \frac{-A_1}{B_1 - A_1}. \tag{21}$$

Similarly,

$$\frac{-A_4}{B_4 - A_4} < \frac{-A_1}{B_1 - A_1}. \tag{22}$$

$$\frac{-A_2}{B_2 - A_2} < \frac{-A_3}{B_3 - A_3}. \tag{23}$$

and

$$\frac{-A_4}{B_4 - A_4} < \frac{-A_3}{B_3 - A_3}. \tag{24}$$

Define

$$\lambda_1 = \min\left(\frac{-A_1}{B_1 - A_1}, \frac{-A_3}{B_3 - A_3}\right).$$

Since $A_1 < 0,\ A_3 < 0,\ B_1 - A_1 > 0,\ B_3 - A_3 > 0$; it follows that $0 < \lambda_1 < 1$. (25)

Define

$$\lambda_2 = \max\left(0, \min\left(\frac{-A_2}{B_2 - A_2}, \frac{-A_4}{B_4 - A_4}\right)\right).$$

Clearly, $\lambda_2 \geq 0$.

In view of (22) - (25), it follows that

$$\lambda_2 < \lambda_1. \tag{26}$$

Define $\lambda_0 = 0.5\ (\lambda_1 + \lambda_2)$. (27)

Clearly, $0 \leq \lambda_0 \leq 1$.

Define $h^e(s) = \lambda_o M_e(s) + (1 - \lambda_o) m_e(s)$. (28)

Claim that

$$h(s) = h^e(s) + s f^o(s) \tag{29}$$

is Hurwitz and lies in the Kharitonov line of $\mathcal{P}_e$. From its very definition, it is obvious that h(s) lies in the Kharitonov line of $\mathcal{P}_e$. To show that h(s) is Hurwitz, it is enough to show that (with $s^2 = -p$)

$$h^e(O_\ell) < 0 \quad \text{and} \quad h^e(O_m) > 0 \quad \text{for} \quad \ell = 1,3 \quad \text{and} \quad m = 2,4.$$

Indeed, $h^e(O_1) = \lambda_0 B_1 + (1-\lambda_0)A_1 < 0$ if and only if

$$\lambda_0 < \frac{-A_1}{B_1 - A_1}$$

which is true by the way λ_O was defined above.

Similarly, one can easily show that

$$h^e(O_3) < 0,\ h^e(O_2) > 0,\ h^e(O_4) > 0. \text{ Hence } h(s) \text{ is Hurwitz.}$$ QED

Next, a set of necessary and sufficient conditions are presented which yield the existence of a Hurwitz polynomial in $\mathcal{P}_e$.

The motivation for the next theorem is to lay the framework for an algorithm to generate a Hurwitz polynomial in $\mathcal{P}_e$, if it exists. In this approach, we keep $M_e(p)$ fixed and vary $m_e(p)$ within the bounds of the interval polytope.

Theorem 3. Suppose for the semi-interval polytope, $\mathcal{P}_e$, $A_1<0$, $A_3<0$, $B_2>0$ and $B_4>0$

The semi-interval polytope $\mathcal{P}_e$ contains a Hurwitz polynomial if and only if the following inequalities have a solution.

(a) $\overline{A}_i < 0, \quad i = 1,3$

where

$$\overline{A}_i = x_9 - x_7 O_i + x_5 O_i^2 - x_3 O_i^3 + x_1 O_i^4, \quad i = 1,3$$

and

$$0 < a_r^- \le x_r \le a_r^+, \ r = 1,3,5,7,9$$

(b)
$$\begin{aligned}(B_\ell - B_m)\overline{\Delta}_9 &+ (O_m B_\ell - O_\ell B_m)\overline{\Delta}_7 \\ &+ (O_m^2 B_\ell - O_\ell^2 B_m)\overline{\Delta}_5 + (O_m^3 B_\ell - O_\ell^3 B_m)\overline{\Delta}_3 \\ &+ (O_m^4 B_\ell - O_\ell^4 B_m)\overline{\Delta}_1 < 0\end{aligned} \tag{30}$$

where $\ell = 1,3;\ m = 2,4$

and

$$\overline{\Delta}_i = \begin{cases} a_i^+ - x_i, & i = 1,5,9 \\ x_i - a_i^-, & i = 3,7 \end{cases}$$

Proof. Suppose $\mathcal{P}_e$ contains a Hurwitz polynomial $h(s)$ whose even part is given by

$$h^e(s) = b_9 + b_7 s^2 + b_5 s^4 + b_3 s^6 + b_1 s^8. \tag{31}$$

Define $x_i = b_i, \quad i = 1,3,5,7,9.$

Define a semi-interval subpolytope $\overline{\mathcal{P}}_e$ of $\mathcal{P}_e$ whose even part is given by

$$[x_9, a_9^+] + [a_7^-, x_7]s^2 + [x_5, a_5^+]s^4 + [a_3^-, x_3]s^6 + [x_1, a_1^+]s^8. \tag{32}$$

and whose odd part is fixed and is the same as that of $\mathcal{P}_e$.

Obviously the K-line of $\overline{\mathcal{P}}_e$ contains a Hurwitz polynomial, viz., $h(s)$ itself and hence, by Theorem 2, all the conditions in (30) are satisfied. Conversely, if the conditions of (30) are satisfied, then by Theorem 2, the K-line of $\overline{\mathcal{P}}_e$ contains a Hurwitz polynomial. Since $\overline{\mathcal{P}}_e$ is contained in $\mathcal{P}_e$, the result follows. QED

Next, we investigate the changes in the expressions for the $H_{\ell m}$'s and the A_i's if the parameters in the lower bound $m_e(p)$ for the even part are varied within the given bounds and the upper bound $M_e(p)$ is kept fixed. So, fix $M_e(p)$ and define

$$\begin{aligned}\overline{m}_e(p) = (a_9^- + \delta_9) - (a_7^+ - \delta_7)p + (a_5^- + \delta_5)p^2 \\ - (a_3^+ - \delta_3)p^3 + (a_1^- + \delta_1)p^4\end{aligned} \tag{33}$$

where

$$0 \le \delta_i \le a_i^+ - a_i^-, \; i = 1,3,5,7,9 \tag{34}$$

Define

$$\begin{aligned}\overline{A}_i &= (a_9^- + \delta_9) - (a_7^+ - \delta_7)O_i + (a_5^- + \delta_5)O_i^2 \\ &\quad -(a_3^+ - \delta_3)O_i^3 + (a_1^- + \delta_1)O_i^4 \\ &= (a_9^- - a_7^+ O_i + a_5^- O_i^2 - a_3^+ O_i^3 + a_1^- O_1^4) \\ &\quad +(\delta_9 + \delta_7 O_i + \delta_5 O_i^2 + \delta_3 O_i^3 + \delta_1 O_1^4) \\ &= A_i + \phi_i, \quad 1 \le i \le 4 \end{aligned} \tag{35}$$

where

$$\phi_i = \delta_9 + \delta_7 O_i + \delta_5 O_i^2 + \delta_3 O_i^3 + \delta_1 O_1^4 \tag{36}$$

$1 \le i \le 4.$

Suppose $\overline{\mathcal{P}}_e$ is the semi-interval subpolytope of $\mathcal{P}_e$ whose lower bound for the even part is $\overline{m}_e(p)$ and the upper bound for the even part is $M_e(p)$ and whose odd part is the same as that of $\mathcal{P}_e$. The corresponding $\overline{H}_{\ell m}$ for $\overline{\mathcal{P}}_e$ are given by

$$\begin{aligned}\overline{H}_{\ell m} &= \overline{A}_\ell B_m - \overline{A}_m B_\ell, \quad \ell = 1,3; \quad m = 2,4 \\ &= (A_\ell + \phi_\ell)B_m - (A_m + \phi_m)B_\ell, \quad \text{by using (35)} \\ &= (A_\ell B_m - A_m B_\ell) + (\phi_\ell B_m - \phi_m B_\ell) \\ &= H_{\ell m} + (\phi_\ell B_m - \phi_m B_\ell) \\ &= H_{\ell m} + (B_m - B_\ell)\delta_9 + (O_\ell B_m - O_m B_\ell)\delta_7 \\ &\quad +(O_\ell^2 B_m - O_m^2 B_\ell)\delta_5 + (O_\ell^3 B_m - O_m^3 B_\ell)\delta_3 \\ &\quad +(O_\ell^4 B_m - O_m^4 B_\ell)\delta_1, \quad \text{by using (36)} \\ &= H_{\ell m} - \phi_{\ell m} \end{aligned} \tag{37}$$

where

$$\begin{aligned}\phi_{\ell m} &= (B_\ell - B_m)\delta_9 + (O_m B_\ell - O_\ell B_m)\delta_7 \\ &\quad +(O_m^2 B_\ell - O_\ell^2 B_m)\delta_5 + (O_m^3 B_\ell - O_\ell^3 B_m)\delta_3 + (O_m^4 B_\ell - O_\ell^4 B_m)\delta_1 \end{aligned} \tag{38}$$

These relations enable us to set up the corresponding linear programming problem such that the existence of a suitable solution of this linear programming problem is equivalent to the existence of a Hurwitz polynomial in $\mathcal{P}_e$.

Suppose, for the semi-interval polytope $\mathcal{P}_e$, $A_1 < 0$, $A_3 < 0$, $B_2 > 0$, $B_4 > 0$. If $H_{\ell m} < 0$ for $l = 1,3$ and $m = 2,4$, then by Theorem 2, the Kharitonov line of $\mathcal{P}_e$ contains a Hurwitz polynomial and we are done. Otherwise, for the sake of illustration, assume that

$$\begin{aligned} H_{12} &= A_1B_2 - A_2B_1 \ge 0 \\ H_{14} &= A_1B_4 - A_4B_1 < 0 \\ H_{32} &= A_3B_2 - A_2B_3 \ge 0 \\ H_{34} &= A_3B_4 - A_4B_3 < 0. \end{aligned} \tag{39}$$

Set up the following linear programming equations in the nonnegative unknowns δ_1, δ_3, δ_5, δ_7, δ_9 and the surplus, or slack nonnegative variables δ_{10}, through δ_{22} as follows:

$$\begin{aligned}&(B_1 - B_2)\delta_9 + (O_2B_1 - O_1B_2)\delta_7 + (O_2^2B_1 - O_1^2B_2)\delta_5 + (O_2^3B_1 - O_1^3B_2)\delta_3 \\ &\quad + (O_2^4B_1 - O_1^4B_2)\delta_1 - \delta_{10} + \delta_{11} = H_{12}\end{aligned} \tag{40}$$

$$-\Big[(B_1 - B_4)\delta_9 + (O_4B_1 - O_1B_4)\delta_7 + (O_4^2B_1 - O_1^2B_4)\delta_5 + (O_4^3B_1 - O_1^3B_4)\delta_3$$
$$+(O_4^4B_1 - O_1^4B_4)\delta_1\Big] + \delta_{12} = -H_{14} \tag{41}$$

$$(B_3 - B_2)\delta_9 + (O_2B_3 - O_3B_2)\delta_7 + (O_2^2B_3 - O_3^2B_2)\delta_5 + (O_2^3B_3 - O_3^3B_2)\delta_3$$
$$+(O_2^4B_3 - O_3^4B_2)\delta_1 - \delta_{13} + \delta_{14} = H_{32} \tag{42}$$

$$-\Big[(B_3 - B_4)\delta_9 + (O_4B_3 - O_3B_4)\delta_7 + (O_4^2B_3 - O_3^2B_4)\delta_5 + (O_4^3B_3 - O_3^3B_4)\delta_3$$
$$+(O_4^4B_3 - O_3^4B_4)\delta_1\Big] + \delta_{15} = -H_{34} \tag{43}$$

$$\delta_9 + O_1\delta_7 + O_1^2\delta_5 + O_1^3\delta_3 + O_1^4\delta_1 + \delta_{16} = -A_1 \tag{44}$$

$$\delta_9 + O_3\delta_7 + O_3^2\delta_5 + O_3^3\delta_3 + O_3^4\delta_1 + \delta_{17} = -A_3 \tag{45}$$

$$\delta_9 + \delta_{18} = \Delta_9 \tag{46}$$

$$\delta_7 + \delta_{19} = \Delta_7 \tag{47}$$

$$\delta_5 + \delta_{20} = \Delta_5 \tag{48}$$

$$\delta_3 + \delta_{21} = \Delta_3 \tag{49}$$

$$\delta_1 + \delta_{22} = \Delta_1 . \tag{50}$$

Equations (40)-(43) are determined by the signs of $H_{\ell m}$ and the relation in (37). Equations (44) and (45) ensure that $\overline{A}_1 < 0$, $\overline{A}_3 < 0$ and the equations (46) - (50) place the constraint on the range of δ_9, δ_7, δ_5, δ_3, δ_1 determined by the dimensions of $\mathcal{P}_e$. Equations (40) - (50) will be called the linear programming equations or the LP equations.

Theorem 5. The semi-interval polytope $\mathcal{P}_e$ contains a Hurwitz polynomial if and only if the linear programming problem as enumerated in equations (40) - (50) has a solution such that

$$\delta_{11} = 0 \quad \text{and} \quad \delta_{14} = 0$$

and (51)

$$\delta_r > 0, \quad r = 10,12,13,15,16,17.$$

Proof. Suppose $\mathcal{P}_e$ contains a Hurwitz polynomial g(s) whose even part is given by $g^e(s) = b_9 + b_7s^2 + b_5s^4 + b_3s^6 + b_1s^8$.

Define

$$\delta_i = \begin{cases} b_i - a_i^-, & i = 1,5,9 \\ a_i^+ - b_i, & i = 3,7 \end{cases} \tag{52}$$

Clearly each $\delta_i \geq 0$, $i = 1,3,5,7,9$.

Since g(s) is Hurwitz, the Kharitonov line of the semi-interval polytope $\overline{\mathcal{P}}_e$ with $M_e(p)$ as the upper bound and $g^e(p)$ as the lower bound for the even part of $\overline{\mathcal{P}}_e$, and with fixed odd part as that of $\mathcal{P}_e$ contains g(s) as a Hurwitz polynomial. Therefore, with obvious notation for $\overline{\mathcal{P}}_e$ and by using Theorem 2, we have

$$g^e(0_i) = \overline{A}_i < 0, \quad \text{for } i = 1,3 \text{ and } \overline{H}_{\ell m} < 0, \ \ell = 1,3; m = 2,4 \tag{53}$$

where $\overline{H}_{\ell m} = (B_\ell - B_m)\overline{\Delta}_9 + (O_mB_\ell - O_\ell B_m)\overline{\Delta}_7$
$+(O_m^2B_\ell - O_\ell^2B_m)\overline{\Delta}_5 + (O_m^3B_\ell - O_\ell^3B_m)\overline{\Delta}_3 + (O_m^4B_\ell - O_\ell^4B_m)\overline{\Delta}_1$

and

$$\overline{\Delta}_i = \begin{cases} a_i^+ - b_i; & i = 1,5,9 \\ b_i - a_i^-; & i = 3,7 \end{cases}$$

Next, we show that the linear programming problem admits of a solution.

First consider (40) for which $H_{12} \geq 0$.

Pick $\delta_{11} = 0$ and $\delta_{10} = \phi_{12} - H_{12}$ (54)

where ϕ_{12} is defined by (38). Clearly δ_{10} and δ_{11} as picked satisfy (40). Using the fact that $\overline{H}_{12} < 0$, and by using (37), it follows that $\phi_{12} > H_{12}$.

Thus, $\delta_{10} > 0$. (55)

Similarly for (42) for which $H_{32} \geq 0$, we pick

$$\delta_{14} = 0, \quad \delta_{13} = \phi_{32} - H_{32}$$

and, as above, it can be shown that $\delta_{13} > 0$ (56)

Now $\overline{H}_{14} < 0$. By using (37), it follows that

$$H_{14} - \phi_{14} < 0. \tag{57}$$

Pick $\delta_{12} = \phi_{14} - H_{14}$, which is positive by using (57). Similarly, since $\overline{H}_{34} < 0$, it follows that $\delta_{15} = \phi_{3,4} - H_{34} > 0$.

Pick $\delta_{16} = -\overline{A}_1$ and $\delta_{17} = -\overline{A}_3$

$$\delta_{16} > 0 \quad \text{and} \quad \delta_{17} > 0 \quad \text{as} \quad \overline{A}_1 < 0 \quad \text{and} \quad \overline{A}_3 < 0.$$

The choice for δ_{18}, δ_{19}, δ_{20}, δ_{21} and δ_{22} is obvious to satisfy (46) - (50). Thus, the linear programming problem has a solution with the constraints as stated.

Conversely. Suppose the linear programming problem as described has a solution. Since $\delta_{11} = 0$ and $\delta_{10} > 0$, it follows from (40) that

$\phi_{12} > H_{12}$.

Since $\overline{H}_{\ell m} = H_{\ell m} - \phi_{\ell m}$ for $\ell = 1,3$; $m = 2,4$, it follows that

$\overline{H}_{12} < 0$.

Similarly, from (42), it follows that

$\overline{H}_{32} < 0$.

From (41), $-\phi_{14} + \delta_{12} = -H_{14}$ which gives

$H_{14} - \phi_{14} = -\delta_{12} < 0$ as $\delta_{12} > 0$,

i.e, $\overline{H}_{14} < 0$ by using (38).

Similarly, $\overline{H}_{34} < 0$.

(44) gives

$$\overline{A}_1 = A_1 + \phi_1 = -\delta_{16} < 0,$$

which yields

$$\overline{A}_1 < 0 \quad \text{by using (35).}$$

Similarly $\overline{A}_3 < 0$.

Thus, we conclude that the semi-interval polytope $\overline{\overline{\mathcal{P}}}_e$ whose even part is given by

$$[a_9^- + \delta_9, a_9^+] + [a_7^-, a_7^+ - \delta_7]s^2 + [a_5^- + \delta_5, a_5^+]s^4$$
$$+ [a_3^-, a_3^+ - \delta_3]s^6 + [a_1^- + \delta_1, a_1^+]s^8$$

and whose odd is fixed and is the same as that of $\mathcal{P}_e$ has the following constraints:

i) $\overline{H}_{12} < 0, \overline{H}_{14} < 0, \overline{H}_{32} < 0, \overline{H}_{34} < 0$

ii) $\overline{A}_1 < 0, \overline{A}_3 < 0, B_2 > 0, B_4 > 0$.

$\therefore$ by Theorem 2, the Kharitonov line of $\overline{\overline{\mathcal{P}}}_e$ contains a Hurwitz polynomial. Since $\overline{\overline{\mathcal{P}}}_e$ is a subset of $\mathcal{P}_e$, it follows that $\mathcal{P}_e$ contains a Hurwitz polynomial. QED

Now we present two numerical examples to illustrate the theory.

Example 1. Consider a semi-interval polytope (whose odd part is fixed) generated by

$$[0.9,1.1]s^{12} + 19.5s^{11} + [20,200]s^{10} + 573s^{9} + [500,2000]s^{8} + 2894.9s^{7} + [1000,8000]s^{6} + 4684s^{5} + [1000,8000]s^{4} + 2894.9s^{3} + [1000,5000]s^{2} + 573s + [10,200] \qquad (58)$$

By keeping the upper bound for the even part fixed, a linear programming problem was established as explained in Theorem 5. A basic feasible solution to the linear programming problem satisfying Theorem 5 was determined which yielded a Hurwitz polynomial given by

$$s^{12} + 19.5s^{11} + 110.7s^{10} + 573s^{9} + 1536.9s^{8} + 2894.9s^{7} + 4535.3s^{6} + 4684s^{5} + 4831.2s^{4} + 2894.9s^{3} + 1859.6s^{2} + 573s + 140.1. \qquad (59)$$

Thus, the given polytope contains the above Hurwitz polynomial.

Summary

A set of necessary and sufficient conditions is proposed which shows whether a semi-interval polytope of polynomials whose odd part is fixed contains a Hurwitz polynomial. These necessary and sufficient conditions are shown to be equivalent to solving a linear programming problem. A numerical example is given to illustrate the proposed conditions.

References

1. Barmish, B. R., New Tools for Robustness of Linear Systems, Macmillan Publishing Company, New York, 1994.
2. Bhattacharyya, S. P.; H. Chapellatt and L. H. Keel, Robust Control, The Parametric Approach, Prentice Hall, 1995.
3. Padmanabhan, P. and C. V. Hollot, “Complete Instability of a Box of Polynomials,” IEEE Transactions on Automatic Control, Vol. 37, pp. 1230-1233, 1992.
4. Pujara, L. R. and Naresh Shanbhag, “Some Stability Theorems for Polygons of Polynomials,” IEEE Transactions on Automatic Control, Vol. 37, pp. 1845-1849, 1992.
5. Pujara, L. R., “On the Interval Polytope Problem,” Proceedings of the 1995 American Control Conference, pp. 966-967.
6. Pujara, L. R., “On the Pseudoboundary of Unstable Polytopes of Polynomials,” IEEE Transactions on Automatic Control, Vol. 41, pp. 1188-1190, 1996.
7. Pujara, L. R., “Necessary and Sufficient Conditions for a Semi-Interval Polytope to Contain a Hurwitz Polynomial”, accepted for the 1997 Allerton Conference, University of Illinois, Urbana, IL, October 1997.

A SUFFICIENT CONDITION FOR THE STABILITY OF MULTI-INPUT MULTI-OUTPUT ARMA MODELS WITH TIME-VARYING COEFFICIENTS

Paul R. Bouthellier
Department of Mathematics
Mount Senario College
1500 College Avenue West
Ladysmith, Wisconsin 54848

Bijoy K. Ghosh
Department of Systems Science and Mathematics
Washington University
Box 1040
St. Louis, MO 63130

ABSTRACT

A sufficiency condition will be derived under which multi-input multi-output (MIMO) autoregressive moving-average (ARMA) models are stable. The algorithm utilizes a previously derived sufficiency condition for the stability of time-varying state-space systems.

1. INTRODUCTION

The problems of ascertaining the stability and stabilizability of linear time-invariant plants by a linear time-invariant output feedback compensator have been among the most important problems in systems theory. However, since many of the problems considered in practice are time-varying, it is important to generalize results well-known in time-invariant systems theory (see [1], [2], [3], [4]) to the time-varying case. Hence in section 2 we shall give a sufficient condition under which a m-input p-output ARMA model with time-varying coefficients is bounded-input bounded-output (BIBO) stable. Our condition will use the recently developed sufficiency condition for ascertaining the stability of time-varying state-space systems [5]. Since many of the results for stabilizing time-varying systems require knowing whether a time-varying MIMO ARMA model is stable [6], [7], our results have immediate applications.

In section 3 we shall give a sufficiency condition for time-varying MIMO ARMA models to be BIBO stable using the notion of time-varying poles of single-input single-output (siso) time-varying difference equations [8]. It will be noted that the sufficiency conditions derived in sections 2 and 3 do not require that the coefficients of the systems considered be periodic or slowly varying. In section 4 concluding remarks will be made.

2. A SUFFICIENCY CONDITION FOR THE STABILITY OF MIMO ARMA MODELS

In this section we will factor the autoregressive (AR) part of an ARMA model into a cascade of 1st order subsystems-each of which corresponds to a time-varying state-space system. The stability of the ARMA model will then be determined by the stability of each of the state-space systems.

We first define the class of systems we shall consider.

Definition 2.1: Let IR$^{i\times j}$ be the space of i×j matrices whose elements belong to IR. A system of the form

$$y(k)+\sum_{i=1}^{l} D_i(k)y(k-i)=\sum_{i=0}^{l} N_i(k)u(k-i) \tag{2.1}$$

where $D_i(k)\in \mathrm{IR}^{p\times p}$ and $N_i(k)\in \mathrm{IR}^{p\times m}$ i=0, 1, 2 ..., l $\forall k\in Z$ is defined to be a m-input p-output ARMA model of lag l. For the sake of notation we shall define $B_l(m,p)$ to be the space of all p×m ARMA models of lag l which satisfy i) $y(i)=\underline{0}\in \mathrm{IR}^p$ and $u(i)=\underline{0}\in \mathrm{IR}^m$ $\forall$ i <0 and ii) $\|D_i(\bullet)\|$ and $\|N_i(\bullet)\|$ i=0, 1, ..., l are uniformly bounded functions of time (where $\|\bullet\|$ denotes the standard Euclidean norm).

All of the results we shall derive in this paper also hold for bounded non-zero initial conditions on y(k)-the proofs, beings similiar, are omitted.

Note: By considering the parameter vector of an ARMA model ($D_1(k)$, $D_2(k)$,..., $D_l(k)$;$N_0(k)$, $N_1(k)$, ..., $N_l(k)$) as a column vector in $IR^{l(p^2+pm)+pm}$ for each time k, time-varying systems of the form (2.1) can be viewed as elements of the product space $\prod_{j=-\infty}^{\infty} IR^{l(p^2+pm)+pm}$. We shall give this space the product topology [9].

To factor (2.1) we will use an operational algebra for time-varying systems which is described as follows: Where $a(k) \in IR^{p\times p}$ $\forall$ $k \in Z$ define the operator $z^{-1}\bullet$ [10], [11] by

$$z^{-1} \bullet a(k) = a(k-1) \tag{2.2}$$

and the operator $z^{-1}\circ$ [10], [11] by

$$z^{-1} \circ a(k) = a(k-1)z^{-1}. \tag{2.3}$$

The following proposition, which lists the key properties of the above operators, may be verified by direct computation and hence the proof is omitted.

Proposition 2.2[12]: Where $IR^{p\times p}[z^{-1}]$ is defined to be the set of power series in z^{-1} with coefficients in $IR^{p\times p}$:

1) If $D_k(z^{-1}) \equiv I + \sum_{i=1}^{l} D_i(k)z^{-i} \in IR^{p\times p}[z^{-1}]$ there exists as unique inverse $D_k^{-1}(z^{-1}) \in R^{p\times p}[z^{-1}]$ such that

$$D_k^{-1}(z^{-1}) \circ D_k(z^{-1}) = D_k(z^{-1}) \circ D_k^{-1}(z^{-1}) \equiv I_{pxp} \qquad \forall k \in Z \tag{2.4}$$

2) Where $A_i(k,z^{-1}) \in IR^{p_i \times p_{i+1}}[z^{-1}]$ i=1, 2, 3 and $y \in IR^{p_4}$, p_i i=1,2,3, and 4 being finite positive integers:

$$[A_1(k,z^{-1}) \circ (A_2(k,z^{-1}) \circ A_3(k,z^{-1}))] \bullet y(k) = A_1(k,z^{-1}) \bullet [A_2(k,z^{-1}) \bullet (A_3(k,z^{-1}) \bullet y(k))] \tag{2.5}$$

3) '$\circ$' is associative.

Using the operators defined above, the system (2.1) may be written as

$$(I + \sum_{i=1}^{l} D_i(k)z^{-i}) \bullet y(k) = (\sum_{i=0}^{l} N_i(k)z^{-i}) \bullet u(k) \tag{2.6}$$

It may be shown [12] that for a generic system (2.6), the AR part may be factored as follows:

$$I + \sum_{i=1}^{l} D_i(k)z^{-i} = (I - \Delta_1(k)z^{-1}) \circ (I - \Delta_2(k)z^{-1}) \circ \cdots \circ (I - \Delta_l(k)z^{-1}) \tag{2.7}$$

where $\Delta_i(k) \in IR^{p\times p}$ i=1, 2, ..., l satisfy

$$D_1(k) = -\sum_{i=1}^{l} \Delta_i(k)$$

$$D_2(k) = \sum_{i=1}^{l-1} \sum_{j>i}^{l} \Delta_i(k)\Delta_j(k-1)$$

$$D_3(k) = -\sum_{i=1}^{l-2} \sum_{j>i}^{l-1} \sum_{n>j}^{l} \Delta_i(k)\Delta_j(k-1)\Delta_n(k-2) \tag{2.8}$$

$$\vdots$$

$$D_l(k) = (-1)^l \Delta_1(k)\Delta_2(k-1)\Delta_3(k-2)\cdots\Delta_l(k-l+1)$$

Note: By generic we shall mean that the set of plants for which the factorization (2.7)-(2.8) can be accomplished may be written as the intersection of a countable number of open and dense sets in $\prod_{j=-\infty}^{\infty} IR^{p^2 l}$ -the product space of the AR part of (2.1)-given the product topology.

Using proposition 2.2, the input-output relationship for system (2.1) may be written as

$$y(k)=[(I-\Delta_l(k)z^{-1})^{-1}\circ(I-\Delta_{l-1}(k)z^{-1})^{-1}\circ\cdots\circ(I-\Delta_1(k)z^{-1})^{-1}]\bullet\left[(\sum_{i=0}^{l}N_i(k)z^{-i})\bullet u(k)\right] \tag{2.9}$$

As each factor $I-\Delta_i(k)z^{-1}$ i=1, 2, ..., l corresponds to a first order state-space system, using proposition 2.2 and (2.7)- (2.9) we get

Theorem 2.3: A sufficient condition for a generic ARMA system of the form (2.1) belonging to the class $B_l(m,p)$ to be BIBO stable is that the l pth-order time-varying state-space systems

$$x_i(k+1)=\Delta_i(k)x_i(k)+v_i(k);\quad x_i(0)=\underline{0}, \tag{2.10}$$

where the $\Delta_i(k)$ i=1, 2,...,l are defined by (2.7), are BIBO stable.

The stability of state-space systems (2.10) was analyzed in [5]. A summary of the results we shall need from [5] is provided below:

For the i$^{\text{th}}$ state-space system given by (2.10), where each component is uniformly reachable over Z [13], [14], i.e. ∀ k ∈Z and j=1, 2,...,p

$$\left|\det(R_{i,j}(k))\right|>\varepsilon>0 \tag{2.11}$$

where

$$R_{i,j}(k)\equiv\left[\,e_j\,\middle|\,\Delta_i(k,k)e_j\,\middle|\,\Delta_i(k,k-1)e_j\,\middle|\cdots\middle|\,\Delta_i(k,k-p+2)e_j\,\right], \tag{2.12}$$

$$\Delta_i(k,k-i)=\Delta_i(k)\Delta_i(k-1)\Delta_i(k-2)\cdots\Delta_i(k-i)\qquad i>0, \tag{2.13}$$

and e_j is the j$^{\text{th}}$ basis vector in IR^p, a sufficient condition for (2.10) to be BIBO stable is given by

Theorem 2.4[5]: Provided each component of the pth-order state-space system (2.10) is uniformly reachable over Z, (2.10) is BIBO stable provided that the p siso difference equations

$$y(k)+\sum_{j=1}^{p}\alpha^i_{p-j}(k)y(k-j)=u_i(k) \tag{2.14}$$

i=1, 2, ...p are BIBO stable. The coefficients $\alpha^i_j(k)$ are found by solving the time-varying version of the Cayley-Hamilton theorem:

$$\Delta_i(k,k-p+1)+\Delta_i(k,k-p+2)\begin{bmatrix}\alpha^1_{p-1}(k-p+1) & 0 & \cdots & 0\\ 0 & \alpha^2_{p-1}(k-p+1) & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \alpha^p_{p-1}(k-p+1)\end{bmatrix}+$$

$$\Delta_i(k,k-p+3)\begin{bmatrix}\alpha^1_{p-2}(k-p+2) & 0 & \cdots & 0\\ 0 & \alpha^2_{p-2}(k-p+2) & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \alpha^p_{p-2}(k-p+2)\end{bmatrix}+\cdots+$$

$$\Delta_i(k)\begin{bmatrix}\alpha^1_1(k-1) & 0 & \cdots & 0\\ 0 & \alpha^2_1(k-1) & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \alpha^p_1(k-1)\end{bmatrix}+\begin{bmatrix}\alpha^1_0(k) & 0 & \cdots & 0\\ 0 & \alpha^2_0(k) & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \alpha^p_0(k)\end{bmatrix}=O_{pxp} \tag{2.15}$$

Proof: We refer the interested reader to [5].

For ease of notation we shall refer to the i$^{\text{th}}$ state-space system (2.10) by S_i and its p corresponding characteristic equations (2.14) by $\{p_{i,j}\}$ j=1, 2,...,p. The stability of siso time-varying difference equations have been studied by many authors, see [8], [11], [12], and the references contained therein. In section 3 we shall give one such stability condition.

Combining the results of thorems 2.3 and 2.4, we see that a generic m-input p-output ARMA model with uniformly bounded time-varying coefficients is stable if each of its l 1$^{\text{st}}$ order factors $\{I-\Delta_i(k)z^{-1}\}_{i=1}^{l}$ corresponds to a state-space system with uniformly reachable components with p corresponding BIBO stable characteristic equations. We now state the main result of this section.

Theorem 2.5: Let S be a generic m-input p-output ARMA model of lag l which satisfies i) $S\in B_l(m,p)$ and ii) each factor $I-\Delta_i(k)z^{-1}$ i=1, 2, ..., l of S's autoregressive part has uniformly reachable components. A sufficient condition for the BIBO stability of S is that each of the l sets of characteristic equations $\{p_{i,j}\}$ i=1, 2,..., l; j=1, 2,...p. consist of BIBO stable polynomials.

Proof: Follows directly from the results of theorem 2.3 and 2.4.

Before an example is considered, two remarks are in order: 1) Note that the problem of studying the stability of a p×m time-varying ARMA model of lag l has been reduced to studying the stability properties of pl siso time-varying difference equations of lag p. 2) It might be tempting to skip the step of factoring the AR part of the ARMA model (2.1) and write it as a lp$^{\text{th}}$ order state-space system in reachable canonical form-then directly apply the result of theorem 2.4 on this one system. This method would leave us however with the problem of having to ascertain the stability of pl siso time-varying difference equations of lag lp-instead of lag p as obtained after factoring. Hence factoring first results in lower order polynomials.

The following example will illustrate the above concepts.

Example 2.6: Consider the 2-input 2-output ARMA model of lag 2 given by

$$y(k)+D_1(k)y(k-1)+D_2(k)y(k-2)=N(k)u(k) \tag{2.16}$$

where

$$D_1(k)=\begin{cases}\begin{bmatrix}-1.1 & -.2\\ -.2 & -1.2\end{bmatrix}, & k\ even\\ \begin{bmatrix}-.5 & -.3\\ -.3 & -.9\end{bmatrix}, & k\ odd\end{cases} \quad \text{and} \quad D_2(k)=\begin{cases}\begin{bmatrix}.11 & .09\\ .08 & .25\end{bmatrix}, & k\ even\\ \begin{bmatrix}.2 & .15\\ .17 & .32\end{bmatrix}, & k\ odd,\end{cases} \tag{2.17}$$

N(k) is a 2×2 matrix with uniformly bounded coefficients, and y(i)=u(i)=$\underline{0}\in \mathrm{IR}^2$ $\forall$ i<0. Using (2.8) the AR part of (2.16) may be written

$$I+D_1(k)z^{-1}+D_2(k)z^{-2}=(I-\Delta_1(k)z^{-1})\circ(I-\Delta_2(k)z^{-1}) \tag{2.18i}$$

where

$$\Delta_1(k)=\begin{cases}\begin{bmatrix}.5 & .1\\ .1 & .6\end{bmatrix}, & k\ even\\ \begin{bmatrix}.3 & .2\\ .2 & .5\end{bmatrix}, & k\ odd\end{cases} \quad \text{and} \quad \Delta_2(k)=\begin{cases}\begin{bmatrix}.6 & .1\\ .1 & .6\end{bmatrix}, & k\ even\\ \begin{bmatrix}.2 & .1\\ .1 & .4\end{bmatrix}, & k\ odd\end{cases} \tag{2.18ii}$$

By theorem 2.3, (2.16)-(2.17) is BIBO stable provided the following state-space systems are BIBO stable:

$$x_1(k+1)=\Delta_1(k)x_1(k)+v_1(k)\,;\quad x_1(0)=\underline{0} \tag{2.19i}$$

and

$$x_2(k+1)=\Delta_2(k)x_2(k)+v_2(k);\quad x_2(0)=\underline{0}\,. \tag{2.19ii}$$

Using (2.15), the p=2 characteristic equations $p_{1,1}$ and $p_{1,2}$ for (2.19i) are given by

$p_{1,1}$: $$y(k)+\alpha_1^1(k)y(k-1)+\alpha_0^1(k)y(k-2)=u_1(k), \tag{2.20i}$$

and

$p_{1,2}$: $$y(k)+\alpha_1^2(k)y(k-1)+\alpha_0^2(k)y(k-2)=u_2(k) \tag{2.20ii}$$

where

$$\alpha_1^1(k)=\alpha_1^2(k)=\begin{cases}-\dfrac{3}{4}, & k\ even\\ -\dfrac{3}{2}, & k\ odd\end{cases}\quad \text{and}\quad \alpha_0^1(k)=\alpha_0^2(k)=\begin{cases}\dfrac{29}{50}, & k\ even\\ \dfrac{11}{200}, & k\ odd.\end{cases} \tag{2.21}$$

(For this example $p_{1,1}=p_{1,2}$.)

The p=2 characteristic equations $p_{2,1}$ and $p_{2,2}$ for (2.19ii) are given by

$p_{2,1}$: $$y(k)+\alpha_1^1(k)y(k-1)+\alpha_0^1(k)y(k-2)=u_1(k) \tag{2.22i}$$

where

$$\left[\alpha_1^1(k),\,\alpha_0^1(k)\right]=\begin{cases}\left[-1,\dfrac{7}{20}\right], & k\ even\\ \left[-\dfrac{4}{5},\dfrac{7}{100}\right], & k\ odd\end{cases} \tag{2.22ii}$$

and

$p_{2,2}$: $$y(k)+\alpha_1^2(k)y(k-1)+\alpha_0^2(k)y(k-2)=u_2(k) \tag{2.23i}$$

where

$$\left[\alpha_1^2(k),\,\alpha_0^2(k)\right]=\begin{cases}\left[-\dfrac{4}{5},\dfrac{7}{20}\right], & k\ even\\ \left[-1,\dfrac{7}{100}\right] & k\ odd\end{cases} \tag{2.23ii}$$

By theorem 2.5, (2.16)-(2.17) is BIBO stable provided the 4 characteristic equations (2.20)-(2.23) are BIBO stable. As it is easy to show that these 4 2-periodic polynomials are stable, the ARMA model (2.16)-(2.17) is BIBO stable.

3. THE TIME-VARYING POLES OF MIMO ARMA MODELS

In this section we shall show that a m-input p-output ARMA model with time-varying coefficients (2.1) is BIBO stable provided its $p^2 l$ time-varying poles satisfy a technical condition.

By theorem 2.5 the p×m lth order system (2.1) is BIBO stable provided its corresponding lp characteristic equations $\{p_{i,j}\}$ i=1,...,l; j=1,...,p are BIBO stable. A typical characteristic equation has the form

$$y(k)+\sum_{j=1}^{p}\alpha_{p-j}(k)y(k-j)=u(k). \tag{3.1}$$

Conditions under which time-varying difference equations (3.1) are stable will now be given. Using the operational algebra defined in section 2, (3.1) may be rewritten as

$$(1+\sum_{j=1}^{p}\alpha_{p-j}(k)z^{-j})\bullet y(k)=u(k). \tag{3.2}$$

Using (2.8) we can factor the AR part of (3.2) into a product of p 1st order factors as follows:

$$1+\sum_{j=1}^{p}\alpha_{p-j}(k)z^{-j}=(1-\delta_1(k)z^{-1})\circ(1-\delta_2(k)z^{-1})\circ\cdots\circ(1-\delta_p(k)z^{-1}) \tag{3.3}$$

where $\delta_i(k)\in C$ $\forall$ $k\in Z$, $i=1,2,\ldots,p$. The $\delta_i(k)$ are called the time-varying poles [8] of the p-th order polynomial (3.3). We now relate the stability of time-varying difference equations to their time-varying poles:

Theorem 3.1 [12]: A sufficient condition for a difference equation (3.1) to be BIBO stable is that its corresponding operator polynomial $1+\sum_{j=1}^{p}\alpha_{p-j}(k)z^{-j}$ admits a factorization of the form (3.3), where each time-varying pole satisfies

$$\sum_{n=0}^{\infty}\prod_{j=0}^{n}|\delta_i(k-j)|<B_i \qquad \forall k\in Z \tag{3.4}$$

$i=1, 2, \ldots, p$ where the B_i, $i=1, 2, \ldots, p$, are finite constants.

Proof: We refer the interested reader to [12].

Note: Condition (3.4) does not require that the time-varying poles lie bounded inside the unit disk at each time instance k in order to insure that (3.1) is BIBO stable. Theorem 3.1 does however generalize the standard time-invariant stability result: A siso time-invariant difference equation is BIBO stable iff all its time-invariant poles lie bounded inside the unit disk.

Combining theorems 2.5 and 3.1 we give the main result of this section:

Theorem 3.2: Let S be a m-input p-output time-varying ARMA model of lag *l* which satisfies the conditions of theorem 2.5. Then a sufficient condition for the BIBO stability of S is for each of its *l*p corresponding time-varying characteristic equations, $p_{i,j}$ $i=1,\ldots,l$; $j=1,\ldots,p$, of lag p to have its poles satisfy condition (3.4).

Proof: Follows directly from theorems 2.5 and 3.1.

Example 3.3: Reconsider the 2×2 ARMA model given by (2.16)-(2.17)

$$y(k)+D_1(k)y(k-1)+D_2(k)y(k-2)=N(k)u(k). \tag{3.5}$$

The 4 siso characteristic equations $p_{1,1}$, $p_{1,2}$, $p_{2,1}$, $p_{2,2}$ of the above equation were defined by (2.20)-(2.23). It can be shown by using (2.8) that each of these 2-periodic time-varying equations have the following sets of time-varying poles:

For $p_{1,1}$ and $p_{1,2}$, the time-varying poles $\delta_1(k)$ and $\delta_2(k)$ are given by

$$\delta_1(k)=\begin{cases}\dfrac{2.2-\sqrt{.2}}{4} & k \text{ even}\\[2ex] \dfrac{.8-\sqrt{.2}}{2} & k \text{ odd}\end{cases} \quad \text{and} \quad \delta_2(k)=\begin{cases}\dfrac{.8+\sqrt{.2}}{4} & k \text{ even}\\[2ex] \dfrac{2.2+\sqrt{.2}}{2} & k \text{ odd}\end{cases} \tag{3.6}$$

For $p_{2,1}$ the time-varying poles are given by

$$\delta_1(k)=\begin{cases}\dfrac{1.08-\sqrt{.0464}}{1.6} & k \text{ even}\\[2ex] \dfrac{2.08-4\sqrt{.0464}}{8} & k \text{ odd}\end{cases} \quad \text{and} \quad \delta_2(k)=\begin{cases}\dfrac{.52+\sqrt{.0464}}{1.6} & k \text{ even}\\[2ex] \dfrac{4.32+4\sqrt{.0464}}{8} & k \text{ odd}\end{cases} \tag{3.7}$$

For $p_{2.2}$ the time-varying poles are given by

$$\delta_1(k)=\begin{cases}\dfrac{1.35-\sqrt{.0725}}{2.5} & k \text{ even}\\[2ex] \dfrac{3.25-5\sqrt{.0725}}{10} & k \text{ odd}\end{cases} \quad \text{and} \quad \delta_2(k)=\begin{cases}\dfrac{.65+\sqrt{.0725}}{2.5} & k \text{ even}\\[2ex] \dfrac{6.75-5\sqrt{.0725}}{10} & k \text{ odd}\end{cases} \tag{3.8}$$

All the poles except $\delta_2(k)$ for $p_{1,1}$ and $p_{1,2}$ lie bounded inside the unit disk for all time instances k. By theorem 3.2, in order to show (2.16)-(2.17) is BIBO stable, it only remains to show that this $\delta_2(k)$ satisfies (3.5). For this we only need show that

$$|\delta_2(even)\delta_2(odd)| < 1 \tag{3.9}$$

as $\delta_2(k)$ is 2-periodic. Since $|\delta_2(even)\delta_2(odd)| = .4127 < 1$ we see that (3.4) is satisfied. Hence, as all the time-varying poles of the characteristic equations satisfy (3.4), the MIMO ARMA model (2.16)-(2.17) is BIBO stable.

4. CONCLUSION

In this paper we have given sufficiency conditions under which MIMO ARMA models with time-varying coefficients are BIBO stable. The condition that all the time-varying poles of the characteristic polynomials satisfy the technical condition given by (3.4) generalizes the well-known time-invariant stability result: If all the time-invariant poles all lie bounded inside the unit disk, the corresponding ARMA model is BIBO stable.

REFERENCES

[1] B. K. Ghosh, *"An Approach to Simultaneous System Design, I, Semialgebraic Geometric Methods,"* SIAM Journal of Control and Optimization, Vol. 24, No. 3, May 1986, pp. 480-496.

[2] B. K. Ghosh, "*An Approach to Simultaneous System Design, II. Nonswitching Gain and Dynamic Feedback Compensation by Algebraic Geometric Methods*," SIAM Journal of Control and Optimization, Vol. 26, No. 4, July 1988, pp. 919-963.

[3] B. K. Ghosh, "*Transcendental and Interpolation Methods in Simultaneous Stabilization and Simultaneous Partial Pole Placement Problems*," SIAM Journal of Control and Optimization, Vol. 24, No. 4, July 1986, pp. 1091-1109.

[4] B. K. Ghosh and C. I. Byrnes, "*Simultaneous Stabilization and Simultaneous Pole-Placement by Non-Switching Dynamic Compensation*," IEEE Trans. on Automatic Control, Vol. AC-28, No. 6, June 1983, pp. 735-741.

[5] P. R. Bouthellier, "*A Sufficient Condition for the Stability of Time-Varying State-Space Systems: A Generalization of the Cayley-Hamilton Theorem*," Proceeding of the 34th Annual Allerton Conference on Communication, Control, and Computing, 1996, pp. 363-371.

[6] B. K. Ghosh and P. R. Bouthellier, "*Simultaneous Coefficient Assignment of Discrete-Time Multi-Input Multi-Output Linear Time-Varying Systems: A New Approach for Compensator Design*," SIAM Journal of Control and Optimization, Vol. 31, No. 6, Nov. 1993, pp. 1438-1461.

[7] B. K. Ghosh and P. R. Bouthellier, "*A Nonrecursive Coefficient Assignment and Stabilization Scheme for Linear Time-Varying Systems*," Proceeding of the 32nd Annual Allerton Conference on Communication, Control, and Computing, 1994, pp. 21-30.

[8] E. W. Kamen, "*The Poles and Zeros of a Linear Time-Varying System*," Linear Algebra and its Applications, 98: 1988, pp. 263-289.

[9] S. Willard, *General Topology*, Addison Wesley, Reading, MA 1970.

[10] E. W. Kamen, P. P. Khargonekar, and K. R. Poolla, "*A Transfer Function Approach to Linear Time-Varying Discrete-Time Systems*," SIAM Journal of Control and Optimization, Vol. 23, No. 4, July 1985, pp. 550-565.

[11] P. R. Bouthellier and B. K. Ghosh, "*Robust Stabilization of Discrete-Time Single-Input Single-Output Linear Time-Varying Systems*," Proceeding of the 27th IEEE Conference on Decision and Control, 1988, Austin, Texas, pp. 39-44.

[12] P. R. Bouthellier, *Analysis and Design of Discrete-Time Linear Time-Varying Systems*, D.Sc. Dissertation, Washington University, St. Louis, MO 1990.

[13] D. S. Evans, "*Finite-Dimensional Realizations of Discrete-Time Weighting Patterns*," SIAM Journal of Applied Mathematics, Vol. 22, No. 1, 1972, pp. 45-67.

[14] L. Weiss, "*Controllability, Realization, and Stability of Discrete-Time Systems*," SIAM Journal of Control and Optimization, Vol. 10, No. 2, 1972, pp. 230-251.

Discretization Schemes of Lyapunov Functional in the Stability Problem of Linear Uncertain Time-Delay Systems

Keqin Gu
Department of Mechanical & Industrial Engineering
Southern Illinois University at Edwardsville
Edwardsville, Illinois 62026, USA

Abstract

The stability problem of linear uncertain time-delay systems is considered using a quadratic Lyapunov functional. The kernel of the functional is chosen as piecewise linear. As a result, the stability condition can be written as a linear matrix inequality, with improved estimate of stability limit over the existing approaches. A number of constraints on the paramters can be introduced to reduce the computational effort needed with some compromised accuracy.

Introduction

Stability and control of uncertain time-delay systems has received significant attentions recently. See, for example, Li and de Souza (1996a, b), Niculescu, et. al. (1996), Phoojaruenchanachai and Furuta (1992), Shen, et. al. (1991), Xu and Liu (1994), and the references therein. Some earlier results are based on a rather simple form of Lyapunov functional, with stability criteria independent of the time-delay. See, for example, Verriest, et. al. (1993), Wang, et. al. (1987), and Xie and de Souza (1993). Although such criteria can be easily reformulated to state feedback control, possibly with H_∞ gain constraints, the results are often overly conservative for practical applications.

Many delay-dependent results have been pursued based on Razumikhin theorem. See for example, Li and de Souza (1996a, b), Niculescu, et. al. (1995, 1996), Xu and Liu (1994). Although these results are usually less conservative than the delay-independent results, they can still be rather conservative. Moreover, there are no obvious ways to obtain less conservative results even if one is willing to commit more computational effort on the problem. Furthermore, most criteria do not reduce to a necessary and sufficient condition when applied to uncertainty-free systems.

Gu (1997) proposed a more general Lyapunov functional approach. The delay interval is divided into N segments, and the kernel is chosen as piecewise linear. For small N, the approach requires rather modest computational effort, and yet results in significant improvements over the existing results. For uncertainty free systems, the analytical results can be approached with large N.

In this article, a more general Lyapunov functional is chosen. As a result, it is necessary to divide the square region $[-r,0] \times [-r,0]$ into $2N^2$ triangular regions, with a piecewise linear kernel. Numerical example shows significant improvement over the approach in Gu (1997) for identical N. A number of computationally less demanding criteria, including the scheme proposed in Gu (1997), can be obtained by introducing additional constraints.

The following notations will be used in this article:

W^T	Transpose of matrix W
$\dot{W}$	Derivative of W with respect to time t
$\dot{W}(\alpha)$	Derivative of W with respect to the argument and evaluated at α
$\partial_1 W(\alpha,\beta)$	Partial derivative of W with respect to α
$\partial_{1+2} W(\alpha,\beta)$	$\partial_1 W(\alpha,\beta) + \partial_2 W(\alpha,\beta)$
$\mathcal{R}$, $\mathcal{R}^k$, $\mathcal{R}^{k\times m}$	The set of real numbers, k-vectors and k by m matrices
r	Time-delay
n	Number of states
n_v	Number of vertices
$\mathcal{C}$	The set of continuous $\mathcal{R}^n$ valued function on $[-r,0]$
N	Number of divisions of the interval $[-r,0]$ in discretization
h	Length of each division $= \frac{r}{N}$
I	Identity matrix of appropriate dimensions
$O_{k\times m}$	kn by mn dimensional zero matrix
I_r	Matrix $[O_{N\times 1}, I]$
I_l	Matrix $[I, O_{N\times 1}]$
I_d	$= I_l - I_r$

Problem statement

Consider the uncertain time-delay system,

$$\dot{x}(t) = A(t)x(t) + B(t)x(t-r), \tag{1}$$

where $x(t) \in \mathcal{R}^n$ is the state; $A(t) \in \mathcal{R}^{n\times n}$ and $B(t) \in \mathcal{R}^{n\times n}$ are uncertain matrices, which are unknown and possibly time-varying, but are known to be bounded by some compact set Ω, i.e.,

$$\begin{aligned} (A(t), B(t)) &\in \Omega \subset R^{n\times 2n}, \\ \text{for all } t &\in (0,\infty). \end{aligned} \tag{2}$$

We will attempt to formulate a practically computable criterion to check the stability of the above system.

For a $t \in (0,\infty)$, define

$$x_t \in \mathcal{C}, \quad x_t(\theta) = x(t+\theta), \qquad \theta \in [-r,0]. \tag{3}$$

Then, the stability of the time-delay system defined in (1) can be studied by choosing a quadratic Lyapunov functional

$$\begin{aligned} V &: \mathcal{C} \mapsto \mathcal{R}, \\ V(\phi) &= \frac{1}{2}\phi^T(0)P\phi(0) \\ &\quad +\phi^T(0)\int_{-r}^{0} Q(\xi)\phi(\xi)d\xi \\ &\quad +\frac{1}{2}\int_{-r}^{0} d\xi \int_{-r}^{0} \phi^T(\xi)R(\xi,\eta)\phi(\eta)d\eta \\ &\quad +\frac{1}{2}\int_{-r}^{0} \phi^T(\xi)S(\xi)\phi(\xi)d\xi, \end{aligned} \tag{4}$$

where

$$P \in \mathcal{R}^{n\times n}, \quad P = P^T > 0, \tag{5}$$

$$\begin{aligned} Q &: [-r,0] \to \mathcal{R}^{n\times n}, & (6)\\ S &: [-r,0] \to \mathcal{R}^{n\times n}, &\\ & \quad S^T(\xi) = S(\xi) > 0, & (7)\\ R &: [-r,0]\times[-r,0] \to \mathcal{R}^{n\times n}, &\\ & \quad R(\eta,\xi) = R^T(\xi,\eta). & (8) \end{aligned}$$

It is well known that (Hale and Lunel, 1993):

Theorem 1 *The system (1) is asymptotically stable if there exists a quadratic Lyapunov functional V of the form (4) such that for some $\epsilon > 0$, and arbitrary $\phi \in \mathcal{C}$*

$$V(\phi) \geq \epsilon\phi^T(0)\phi(0), \tag{9}$$

and its derivative along the solution of (1) satisfies

$$\dot{V}(\phi) \leq -\epsilon\phi^T(0)\phi(0), \tag{10}$$

where

$$\dot{V}(\phi) \triangleq \frac{d}{dt}V(x_t)|_{x_t=\phi}.$$

Using (1), one obtains

$$\begin{aligned} &\dot{V}(\phi) = \\ &\phi^T(0)P[A(t)\phi(0) + B(t)\phi(-r)] \\ &+[A(t)\phi(0) + B(t)\phi(-r)]^T \int_{-r}^{0} Q(\xi)\phi(\xi)d\xi \\ &+\phi^T(0)\int_{-r}^{0} Q(\xi)\dot{\phi}(\xi)d\xi \\ &+\int_{-r}^{0} d\xi \int_{-r}^{0} \phi^T(\xi)R(\xi,\eta)\dot{\phi}(\eta)d\eta \\ &+\int_{-r}^{0} \phi^T(\xi)S(\xi)\dot{\phi}(\xi)d\xi. \end{aligned}$$

Through integration by parts, recognizing (8), it is not difficult to see that

$$\begin{aligned} &\dot{V}(\phi) = \\ &-\frac{1}{2}\phi^T(0)[-PA - A^TP - Q(0) \\ &\quad -Q^T(0) - S(0)]\phi(0) \\ &-\frac{1}{2}\phi^T(-r)S(-r)\phi(-r) \\ &-\frac{1}{2}\int_{-r}^{0} \phi^T(\xi)\dot{S}(\xi)\phi(\xi)d\xi \\ &-\frac{1}{2}\int_{-r}^{0} d\xi \int_{-r}^{0} \phi^T(\xi)\partial_{1+2}R(\xi,\eta)\phi(\eta)d\eta \\ &+\phi^T(0)[PB - Q(-r)]\phi(-r) \\ &+\phi^T(0)\int_{-r}^{0} [A^TQ(\xi) - \dot{Q}(\xi) \\ &\quad +R^T(\xi,0)]\phi(\xi)d\xi \\ &+\phi^T(0)\int_{-r}^{0} [B^TQ(\xi) \\ &\quad -R^T(\xi,-r)]\phi(\xi)d\xi. \end{aligned} \tag{11}$$

Discretization

Equations (4) and (11) indicate that the Lyapunov functional and its derivative are both quadratic functionals. We will choose Q, R and S to be piecewise linear, which will enable us to write conditions (9) and (10) as linear matrix inequalities. Specifically, the interval

$$\mathcal{I} = [-r,0]$$

is divided into N small intervals,

$$\mathcal{I}_i = [\theta_{i-1},\theta_i], \quad i = 1,2,...,N,$$

where

$$\begin{aligned} \theta_i &= -r + ih, \\ h &= r/N. \end{aligned}$$

Similarly, the region

$$\mathcal{S} = [-r,0]\times[-r,0]$$

is divided into $N \times N$ small square regions

$$\mathcal{S}_{ij} = [\theta_{i-1},\theta_i]\times[\theta_{j-1},\theta_j], \quad i,j = 1,2,...,N.$$

Each small square is further divided into two triangular region

$$\begin{aligned} \mathcal{T}_{ij}^{u} &= \left\{(\theta_{i-1}+\alpha h, \theta_{j-1}+\beta h) \,\middle|\, \begin{array}{l} 0 \leq \beta \leq 1, \\ 0 \leq \alpha \leq \beta \end{array} \right\}, \\ \mathcal{T}_{ij}^{l} &= \left\{(\theta_{i-1}+\alpha h, \theta_{j-1}+\beta h) \,\middle|\, \begin{array}{l} 0 \leq \alpha \leq 1, \\ 0 \leq \beta \leq \alpha \end{array} \right\}. \end{aligned}$$

Choose Q and S to be continuous in $\mathcal{I}$ and linear within each $\mathcal{I}_i$, and R to be continuous in $\mathcal{S}$ and linear within each $\mathcal{T}_{ij}^u$ and $\mathcal{T}_{ij}^l$. In other words, for $0 \leq \alpha \leq 1$

$$\begin{aligned} Q(\theta_{i-1}+\alpha h) &= Q^i(\alpha) && (12)\\ &= (1-\alpha)Q_{i-1}+\alpha Q_i \\ S(\theta_{i-1}+\alpha h) &= S^i(\alpha) && (13)\\ &= (1-\alpha)S_{i-1}+\alpha S_i \end{aligned}$$

and for $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$,

$$R(\theta_{i-1}+\alpha h, \theta_{j-1}+\beta h) = R^{ij}(\alpha,\beta) \qquad (14)$$

$$= \begin{cases} (1-\alpha)R_{i-1,j-1}+\beta R_{ij}+(\alpha-\beta)R_{i,j-1}, & \alpha \geq \beta \\ (1-\beta)R_{i-1,j-1}+\alpha R_{ij}+(\beta-\alpha)R_{i-1,j}, & \alpha < \beta \end{cases}$$

It is also convenient to write

$$\phi^i(\alpha) = \phi(\theta_{i-1}+\alpha h). \qquad (15)$$

It is easy to see that

$$\begin{aligned} S_i &= S_i^T > 0, \\ R_{ij} &= R_{ji}^T. \end{aligned}$$

Lyapunov functional condition

With the choice of piecewise linear functions, we will write condition (9) in linear matrix inequality. We have

Theorem 2 *With the piecewise linear Q, S and R chosen, the Lyapunov functional satisfies (9) if there exist $n \times n$ matrices $W_i, i = 0, 1, ..., N$ and $U = U^T$ such that*

$$\begin{pmatrix} U & \tilde{W} \\ \tilde{W}^T & \tilde{R} \end{pmatrix} \geq 0, \qquad (16)$$

and

$$\begin{pmatrix} (P-U)/h & \tilde{Q}-\tilde{W} \\ (\tilde{Q}-\tilde{W})^T & \hat{S} \end{pmatrix} > 0, \qquad (17)$$

where

$$\tilde{R} = \begin{pmatrix} R_{00} & R_{01} & \dots & R_{0N} \\ R_{10} & R_{11} & \dots & R_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ R_{N0} & R_{N1} & \dots & R_{NN} \end{pmatrix} \qquad (18)$$

$$\begin{aligned} \tilde{W} &= (W_0, W_1, ..., W_N), && (19)\\ \tilde{Q} &= (Q_0, Q_1, ..., Q_N), && (20)\\ \hat{S} &= \mathrm{diag}(2S_0, S_1, S_2, ..., \\ & \qquad S_{N-2}, S_{N-1}, 2S_N). && (21) \end{aligned}$$

Proof. Define

$$W(\theta_{i-1}+\alpha h) = W^i(\alpha) = (1-\alpha)W_{i-1}+\alpha W_i. \qquad (22)$$

Then, we can write

$$V(\phi) = V_1(\phi) + V_2(\phi),$$

where

$$\begin{aligned} V_1(\phi) &= \frac{1}{2}\phi^T(0)U\phi(0) \\ &\quad +\phi^T(0)\int_{-r}^0 W(\xi)\phi(\xi)d\xi \\ &\quad +\frac{1}{2}\int_{-r}^0 d\xi \int_{-r}^0 \phi^T(\xi)R(\xi,\eta)\phi(\eta)d\eta, \end{aligned}$$

and

$$\begin{aligned} V_2(\phi) &= \frac{1}{2}\phi^T(0)(P-U)\phi(0) \\ &\quad +\phi^T(0)\int_{-r}^0 [Q(\xi)-W(\xi)]\phi(\xi)d\xi \\ &\quad +\frac{1}{2}\int_{-r}^0 \phi^T(\xi)S(\xi)\phi(\xi)d\xi. \end{aligned}$$

It is sufficient to prove:

(i) (16) guarantees

$$V_1(\phi) \geq 0, \qquad (23)$$

(ii) (17) guarantees

$$V_2(\phi) \geq \epsilon\phi^T(0)\phi(0) \qquad (24)$$

for some sufficiently small $\epsilon > 0$.

To prove (i), use (22), (14) and (15), with a change of order of integration between α and β, we can write

$$V_1(\phi) = \frac{1}{2}\phi^T(0)U\phi(0)$$
$$+h\phi^T(0)\sum_{i=1}^{N}\int_0^1 W^i(\alpha)\phi^i(\alpha)d\alpha$$
$$+h^2\sum_{i=1}^{N}\sum_{j=1}^{N}\int_0^1 d\alpha\int_0^\alpha \phi^{iT}(\alpha)R^{ij}(\alpha,\beta)\phi^j(\beta)d\beta$$

Introduce a change of variable

$$\psi^i(\alpha) = \int_0^\alpha \phi^i(\tau)d\tau$$

or

$$\phi^i(\alpha) = \dot{\psi}^i(\alpha), \quad \psi^i(0) = 0,$$

with integration by parts, it can be seen that

$$V_1(\phi) = \frac{1}{2}\phi^T(0)U\phi(0)$$
$$+h\sum_{i=1}^{N}\Big[\phi^T(0)W_i\psi^i(1)$$
$$-\phi^T(0)\int_0^1(W_i - W_{i-1})\psi^i(\alpha)d\alpha\Big]$$
$$+h^2\sum_{i=1}^{N}\sum_{j=1}^{N}\Big[\psi^{iT}(1)R_{ij}\psi^j(1)$$
$$-\psi^{iT}(1)(R_{ij} - R_{i,j-1})\int_0^1\psi^j(\beta)d\beta\Big]$$
$$-\int_0^1\frac{h^2}{2}\frac{d}{dt}\sum_{i=1}^{N}\sum_{j=1}^{N}\psi^{iT}(\alpha)\Big[(1-\alpha)R_{i-1,j-1}$$
$$+\alpha R_{ij}\psi^j(\alpha)\Big]d\alpha$$
$$-\int_0^1\frac{h^2}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\psi^{iT}(\alpha)$$
$$(-R_{i-1,j-1} + R_{ij})\,\psi^j(\alpha)d\alpha$$
$$-h^2\int\sum_{i=1}^{N}\sum_{j=1}^{N}\psi^{iT}(\alpha)$$
$$(-R_{i-1,j-1} + R_{i,j-1})\,\psi^j(\alpha)d\alpha,$$

or

$$V_1(\phi) = \frac{1}{2}\phi^T(0)U\phi(0)$$
$$+h\sum_{i=1}^{N}\Big[\phi^T(0)W_i\psi^i(1)$$
$$-\phi^T(0)\int_0^1(W_i - W_{i-1})\psi^i(\alpha)d\alpha\Big]$$
$$+h^2\sum_{i=1}^{N}\sum_{j=1}^{N}\Big[\psi^{iT}(1)R_{ij}\psi^j(1)$$
$$-\psi^{iT}(1)(R_{ij} - R_{i,j-1})\int_0^1\psi^j(\beta)d\beta\Big]$$
$$+\int_0^1\frac{h^2}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\psi^{iT}(\alpha)\Big[R_{ij} - R_{i,j-1}$$
$$-R_{i-1,j} + R_{i-1,j-1}\psi^j(\alpha)\Big]d\alpha.$$

Let

$$\tilde{\psi}(\alpha) = (\psi^{1T}(\alpha), \psi^{2T}(\alpha), ..., \psi^{NT}(\alpha))^T,$$

then, the above can be written as

$$V_1(\phi) = \frac{1}{2}\phi^T(0)U\phi(0)$$
$$+h\phi^T(0)\tilde{W}I_r^T\tilde{\psi}(1) - h\phi^T(0)\tilde{W}I_d^T\int_0^1\tilde{\psi}(\alpha)d\alpha$$
$$+\frac{h^2}{2}\tilde{\psi}^T(1)I_r\tilde{R}I_r^T\tilde{\psi}(1)$$
$$-h^2\tilde{\psi}^T(1)I_r\tilde{R}I_d^T\int_0^1\psi^j(\beta)d\beta$$
$$+\frac{h^2}{2}\int_0^1\tilde{\psi}^T(\alpha)I_d\tilde{R}I_d^T\psi^j(\alpha)d\alpha,$$

or

$$V_1(\phi) = \frac{1}{2}\int_0^1(\phi^T(0), h[\psi^T(1)I_r - \psi^T(\alpha)I_d])$$
$$\begin{pmatrix}\tilde{U} & \tilde{W}\\ \tilde{W}^T & \tilde{R}\end{pmatrix}\begin{pmatrix}\phi(0)\\ h[I_r^T\psi(1) - I_d^T\psi(\alpha)]\end{pmatrix}d\alpha.$$

Therefore, (i) is valid.

To prove (ii), write

$$V_2(\phi) = \frac{1}{2}\int_{-r}^1(\phi^T(0), \phi^T(\xi))$$
$$\begin{pmatrix}\frac{1}{r}[P-U] + Z(\xi) & Q(\xi) - W(\xi)\\ [Q(\xi) - W(\xi)]^T & S(\xi)\end{pmatrix}$$
$$\begin{pmatrix}\phi(0)\\ \phi(\xi)\end{pmatrix}d\xi$$

for any $Z(\xi)$ satisfying

$$\int_{-r}^{1} Z(\xi)d\xi = 0. \qquad (25)$$

Therefore, (24) is satisfied if we can find a Z satisfying (25) such that

$$\begin{pmatrix} \frac{1}{r}[P-U]+Z(\xi) & Q(\xi)-W(\xi) \\ [Q(\xi)-W(\xi)]^T & S(\xi) \end{pmatrix} > 0, (26)$$
$$\text{for } -r \leq \xi \leq 0.$$

Choose $Z(\xi)$ to be piecewise linear,

$$Z(\theta_{i-1}+\alpha h) = (1-\alpha)Z_{i-1}+\alpha Z_i, \quad 0 \leq \alpha \leq 1. \qquad (27)$$

Then, the inequality in (26) is satisfied for all $-r \leq \xi \leq 0$ if it is satisfied for $\xi = \theta_i$, $i = 0, 1, ..., N$, in other words

$$\begin{pmatrix} \frac{1}{r}[P-U]+Z_i & Q_i-W_i \\ [Q_i-W_i]^T & S_i \end{pmatrix} > 0, \qquad (28)$$
$$\text{for } i = 0, 1, ..., N,$$

and (25) is equivalent to

$$\frac{1}{2}(Z_0+Z_N)+\sum_{i=1}^{N-1} Z_i = 0 \qquad (29)$$

Therefore, we need only to prove that (17) implies (in fact, is equivalent to) the existence of Z_i such that (29) and (28) are satisfied. Let

$$\Gamma_i = \frac{1}{r}(P-U)-(Q_i-V_i)S_i^{-1}(Q_i-V_i)^T,$$

and

$$\Gamma = \frac{1}{2}(\Gamma_0+\Gamma_N)+\sum_{i=1}^{N-1}\Gamma_i.$$

Then, by Shur's complement, (17) implies

$$\Gamma > 0.$$

Choose sufficiently small $\delta_i > 0$, and let

$$\begin{aligned} Z_i &= \delta_i I - \Gamma_i, \quad i = 1, 2, ..., N, \\ Z_0 &= 2\left[\Gamma - \left(\frac{1}{2}\delta_N + \sum_{i=1}^{N-1}\delta_i\right)I\right] - \Gamma_0 \end{aligned}$$

Then,

$$\Gamma_i + Z_i > 0, \quad i = 0, 1, ..., N.$$

But the above implies (28) by Shur's complement. (29) can be verified by direct calculation. ■

Lyapunov derivative condition

Similar to the last section, we can write (10) in a linear matrix inequality, as is stated in the following Theorem.

Theorem 3 *With the piecewise linear Q, S and R chosen, (10) is satisfied if there exist $n \times n$ matrices*

$$X_{ij}, \quad i = 1, 2; \quad j = 1, 2, ..., N,$$

and

$$T_{ij}^T = T_{ji}, Y_{ij}^T = Y_{ji}, \quad i, j = 1, 2,$$

such that

$$\begin{pmatrix} T_{11} & T_{12} & X_1 \\ & T_{22} & X_2 \\ \textit{symmetric} & & R_d \end{pmatrix} \geq 0, \qquad (30)$$

$$\begin{pmatrix} \Delta_{11}-T_{11}+Y_{11} & -\Delta_{12}-T_{12}+Y_{12} & -D_1^0-X_1 \\ & \Delta_{22}-T_{22}+Y_{22} & -D_2^0-X2 \\ \textit{symmetric} & & \frac{1}{h}S_d \end{pmatrix} > 0, \qquad (31)$$

$$\begin{pmatrix} \Delta_{11}-T_{11}-Y_{11} & -\Delta_{12}-T_{12}-Y_{12} & -D_1^1-X_1 \\ & \Delta_{22}-T_{22}-Y_{22} & -D_2^1-X2 \\ \textit{symmetric} & & \frac{1}{h}S_d \end{pmatrix} > 0, \qquad (32)$$

for all $(A, B) \in \Omega$, where

$$\begin{aligned} \Delta_{11} &= -PA-A^TP-Q_N-Q_N^T-S_N, \qquad (33) \\ \Delta_{12} &= PB-Q_0, \\ \Delta_{22} &= S_0, \end{aligned}$$

$$\begin{aligned}
S_d &= \operatorname{diag}(S_{d1}, S_{d2}, ..., S_{dN}),\\
S_{di} &= \frac{1}{h}(S_i - S_{i-1}),\\
R_d &= \begin{pmatrix} R_{d11} & R_{d12} & \dots & R_{d1N}\\ R_{d21} & R_{d22} & \dots & R_{d2N}\\ \vdots & \vdots & \ddots & \vdots\\ R_{dN1} & R_{dN2} & \dots & R_{dNN}\end{pmatrix},\\
R_{dij} &= \frac{1}{h}(R_{ij} - R_{i-1,j-1}),\\
D_j^k &= (D_{j1}^k, D_{j2}^k, ..., D_{jN}^k),\\
D_{1i}^k &= A^T Q_{i-1+k} - \frac{1}{h}(Q_i - Q_{i-1}) + R_{i-1+k,N}^T,\\
D_{2i}^k &= B^T Q_{i-1+k} - R_{i-1+k,0}^T,\\
X_i &= (X_{i1}, X_{i2}, ..., X_{iN}). \qquad (34)
\end{aligned}$$

Proof. Direct calculation yields

$$\begin{aligned}
&\dot V(\phi) =\\
&-\frac{1}{2}\phi^T(0)\Delta_{11}\phi(0) - \frac{1}{2}\phi^T(-r)\Delta_{22}\phi(-r)\\
&\quad +\phi^T(0)\Delta_{12}\phi(-r)\\
&-\frac{h}{2}\sum_{i=1}^{N}\int_0^1 \phi^{iT}(\alpha)S_{di}\phi^i(\alpha)d\alpha\\
&-\frac{h^2}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\int_0^1 \phi^i(\alpha)d\alpha\right)^T R_{dij}\left(\int_0^1 \phi^i(\alpha)d\alpha\right)\\
&+h\phi^T(0)\sum_{i=1}^{N}\int_0^1\left[(1-\alpha)D_{1i}^0 + \alpha D_{1i}^1\right]\phi^i(\alpha)d\alpha\\
&+h\phi^T(-r)\sum_{i=1}^{N}\int_0^1\left[(1-\alpha)D_{2i}^0 + \alpha D_{2i}^1\right]\phi^i(\alpha)d\alpha,
\end{aligned}$$

or

$$\begin{aligned}
&\dot V(\phi) = -\frac{1}{2}\int_0^1 (\phi^T(0), \phi^T(-r), h\tilde\phi^T(\alpha))\\
&\begin{pmatrix} \Delta_{11} & -\Delta_{12} & -(1-\alpha)D_1^0 - \alpha D_1^1\\ & \Delta_{22} & -(1-\alpha)D_2^0 - \alpha D_2^1\\ \text{symmetric} & & \frac{1}{h}S_d\end{pmatrix}\\
&\begin{pmatrix}\phi(0)\\ \phi(-r)\\ h\tilde\phi(\alpha)\end{pmatrix} d\alpha\\
&-\frac{h^2}{2}\left(\int_0^1 \tilde\phi(\alpha)d\alpha\right)^T R_d\left(\int_0^1 \tilde\phi(\alpha)d\alpha\right).
\end{aligned}$$

It can be further written as

$$\begin{aligned}
&\dot V(\phi) = -\frac{1}{2}\int_0^1 (\phi^T(0), \phi^T(-r), h\tilde\phi^T(\alpha))\\
&\begin{pmatrix} \Delta_{11} - T_{11} + (1-2\alpha)Y_{11} & -\Delta_{12} - T_{12} + (1-2\alpha)Y_{12} & -(1-\alpha)D_1^0 - \alpha D_1^1 - X_1\\ & \Delta_{22} - T_{22} + (1-2\alpha)Y_{22} & -(1-\alpha)D_2^0 - \alpha D_2^1 - X_2\\ \text{Symmetric} & & \frac{1}{h}S_d\end{pmatrix}\begin{pmatrix}\phi(0)\\ \phi(-r)\\ h\tilde\phi(\alpha)\end{pmatrix} d\alpha\\
&-\frac{1}{2}(\phi^T(0), \phi^T(-r), h\int_0^1 \tilde\phi^T(\alpha)d\alpha)\\
&\begin{pmatrix} T_{11} & T_{12} & X_1\\ & T_{22} & X_2\\ \text{symmetric} & & R_d\end{pmatrix}\begin{pmatrix}\phi(0)\\ \phi(-r)\\ h\int_0^1 \tilde\phi(\alpha)d\alpha\end{pmatrix}.
\end{aligned}$$

Therefore, (10) is satisfied if

$$\begin{pmatrix} T_{11} & T_{12} & X_1\\ & T_{22} & X_2\\ \text{symmetric} & & R_d\end{pmatrix} \geq 0,$$

and

$$\begin{pmatrix} \Delta_{11} - T_{11} + (1-2\alpha)Y_{11} & -\Delta_{12} - T_{12} + (1-2\alpha)Y_{12} & -(1-\alpha)D_1^0 - \alpha D_1^1 - X_1\\ & \Delta_{22} - T_{22} + (1-2\alpha)Y_{22} & -(1-\alpha)D_2^0 - \alpha D_2^1 - X_2\\ \text{Symmetric} & & \frac{1}{h}S_d\end{pmatrix} > 0,$$

for $0 \leq \alpha \leq 1$ and $(A, B) \in \Omega$. But this is equivalent to (30),(31) and 32) since the above depend on α linearly, and therefore, needs only to be valid for $\alpha = 0$ and 1. ■

Computational Consideration

The LMI obtained needs to hold for all the possible uncertain system matrix. A relatively general block diagonal uncertainty will be discussed in Gu (1998). Here, we assume the uncertainty set is polytopic

$$\Omega = \mathrm{co}\{(A_i, B_i \mid i = 1, 2, ..., n_v\}. \qquad (35)$$

Then, one only needs to check the satisfaction of the LMI at the vertices. Therefore, we have

Theorem 4 *The system (1), (2) with polytopic uncertainty (35) is stable if there exist real $n \times n$ matrices*

$$\begin{aligned}
&P = P^T, \\
&Q_i, \quad i = 0, 1, ..., N, \\
&R_{ij} = R_{ji}^T, \quad i, j = 0, 1, ..., N, \\
&S_i = S_i^T, \quad i = 0, 1, ..., N, \\
&U = U^T, \\
&W_i, \quad i = 1, 2, ..., N, \\
&X_{ij}, \quad i = 1, 2; \; j = 1, 2, ..., N, \\
&T_{ij} = T_{ij}^T, Y_{ij} = Y_{ji}^T \quad i, j = 1, 2,
\end{aligned}$$

such that (16), (17), (30), are satisfied, and (31) and (32) are satisfied for all

$$(A, B) = (A_i, B_i), \quad i = 1, 2, ..., n_v,$$

where additional notations are defined in (18) to (21) and (33) to (34).

The above LMI typically involve a large number of parameters. Some reduction of parameters are possible by introducing constraints. For example, one can choose

$$\begin{aligned}
W_i &= Q_i, \\
R_{ij} &= R_{i-j}, \\
T_{ij} &= 0, \\
X_{ij} &= 0.
\end{aligned}$$

Then, one obtains

Corollary 5 *The system (1), (2) with polytopic uncertainty (35) is stable if there exist real $n \times n$ matrices*

$$\begin{aligned}
&P = P^T, \\
&Q_i, \quad i = 0, 1, ..., N, \\
&R_i = R_{-i}^T, \quad i = 0, 1, ..., N, \\
&S_i = S_i^T, \quad i = 0, 1, ..., N, \\
&U = U^T, \\
&Y_{ij} = Y_{ji}^T \quad i, j = 1, 2,
\end{aligned}$$

such that

$$\begin{pmatrix} U & \tilde{Q} \\ \tilde{Q}^T & \tilde{R} \end{pmatrix} \geq 0,$$

$$P > U,$$

$$S_i > 0,$$

and for all $(A, B) = (A_i, B_i), i = 1, 2, ..., n_v$,

$$\begin{pmatrix} \Delta_{11} + Y_{11} & -\Delta_{12} + Y_{12} & -D_1^0 \\ & \Delta_{22} + Y_{22} & -D_2^0 \\ \mathit{symmetric} & & \frac{1}{h} S_d \end{pmatrix} > 0,$$

$$\begin{pmatrix} \Delta_{11} - Y_{11} & -\Delta_{12} - Y_{12} & -D_1^1 \\ & \Delta_{22} - Y_{22} & -D_2^1 \\ \mathit{symmetric} & & \frac{1}{h} S_d \end{pmatrix} > 0,$$

where

$$\begin{aligned}
\Delta_{11} &= -PA - A^T P - Q_N - Q_N^T - S_N, \\
\Delta_{12} &= PB - Q_0, \\
\Delta_{22} &= S_0, \\
S_d &= \mathrm{diag}(S_{d1}, S_{d2}, ..., S_{dN}), \\
S_{di} &= \frac{1}{h}(S_i - S_{i-1}), \\
D_j^k &= (D_{j1}^k, D_{j2}^k, ..., D_{jN}^k), \\
D_{1i}^k &= A^T Q_{i-1+k} - \frac{1}{h}(Q_i - Q_{i-1}) + R_{i-1+k-N}^T, \\
D_{2i}^k &= B^T Q_{i-1+k} - R_{i-1+k}^T.
\end{aligned}$$

The above is roughly equivalent to that proposed in Gu (1997).

Numerical example

To illustrate the effectiveness of the approach, two numerical examples are presented.

Example 6 *This example is from Li and de Souza (1996a) and Niculescu, et. al. (1995). This is a linear time invariant time delay system,*

$$\dot{x}(t) = \begin{pmatrix} -2 & 0 \\ 0 & -0.9 \end{pmatrix} x(t) + \begin{pmatrix} -1 & 0 \\ -1 & -1 \end{pmatrix} x(t-r). \tag{36}$$

The maximum time-delay for stability, $r_{\max}$, can be calculated analytically as 6.17. Theorem 4 and Corollary 5 have been used to estimate $r_{\max}$ by solving the linear matrix inequalities with different N. The results are listed in the following.

N	1	2	5	10	20
Theorem 4	*5.30*	*5.74*	*6.07*	*6.14*	*6.16*
Corollary 5	*5.30*	*5.74*	*5.99*	*6.06*	*6.10*

As expected, the results using Theorem 4 approaches the analytical results faster, at a cost of more computation requirement. The results here are much less conservative than the results as surveyed in Li and de Souza (1996a).

The next example illustrate the application of the approach to a polytopic uncertain system.

Example 7 *Consider the following uncertain system,*

$$\begin{aligned} \dot{x}(t) &= \begin{pmatrix} -2+\rho(t) & \rho(t) \\ \rho(t) & -0.9+\rho(t) \end{pmatrix} x(t) \\ &\quad + \begin{pmatrix} -1+\rho(t) & 0 \\ -1 & -1-\rho(t) \end{pmatrix} x(t-r), \end{aligned}$$

where

$$|\rho(t)| \leq 0.1.$$

It can be modeled as a polytopic system, with $n_v = 2$, and

$$A_1 = \begin{pmatrix} -2.1 & -0.1 \\ -0.1 & -1.0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} -1.1 & 0 \\ -1 & -0.9 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} -1.9 & 0.1 \\ 0.1 & -0.8 \end{pmatrix}, \quad B_2 = \begin{pmatrix} -0.9 & 0 \\ -1 & -1.1 \end{pmatrix}.$$

The maximum delay for stability $r_{\max}$ is listed in the following table

N	*1*	*3*	*5*	*10*
Theorem 4	*2.37*	*2.58*	*2.62*	*2.64*
Corollary 5	*2.36*	*2.55*	*2.59*	*2.61*

Conclusions

A new LMI approach for uncertain linear time-delay system stability problem has been formulated. Different simplifications are possible with a cost of convergence speed.

References

Boyd, S. and Yang, Q., 1989, Structured and simultaneous Lyapunov functions for system stability problems, *International Journal of Control*, **49**, 2215–2240.

Boyd, S., et. al., 1994, *Linear Matrix Inequalities in System and Control Theory* (Philadelphia: SIAM).

Gahinet, P., et. al., 1995, *LMI Control Toolbox for use with MATLAB* (Natick, MA: MathWorks).

Gu, K., 1995, The algorithm for studying the trade-offs between H_∞ attenuation and uncertainty tolerance, *International Journal of Control*, **61**, 823-836.

Gu, K., 1997a, Constrained LMI set in the stability problem of linear uncertain time-delay systems, *1997 American Control Conference*, Albuquerque, NM, June 3–8.

Gu, K., 1998, Stability of time-delay systems under block-diagonal uncertainty, to be submitted to *1998 American Control Conference.*

Hale, J. and Lunel, S. M. V., 1993, *Introduction to Functional Differential Equations* (New York: Springer-Verlag).

Li, X. and de Souza, C. E., 1996a, Criteria for robust stability of uncertain linear systems with time-varying state delays, *13th World Congress of IFAC,* San Francisco, CA, USA, June 30–July 5, H, 137–142.

Li, X. and de Souza, C. E., 1996b, Robust stabilization and H_∞ control of uncertain linear time-delay systems, *13th World Congress of IFAC,* San Francisco, CA, USA, June 30–July 5, H, 113–118.

Nesterov, Y. and Nemirovskii, A., 1994, *Interior-Point Polynomial Algorithms in Convex Programming* (Philadelphia: SIAM).

Niculescu, S. I., Dugard, L. and Dion, J. M., 1995, Stabilité et stabilisation robustes des systèmes à retard, *Proc. Journées Robustesse,* Toulouse, France.

Niculescu, S.-I., et. al., 1996, Robust stabilization for uncertain time-delay systems containing saturating actuators, *IEEE Transactions on Automatic Control,* **41**, 742–747.

Phoojaruenchanachai, S. and Furuta, K. 1992, Memoryless stabilization of uncertain linear systems including time-varying state delays, *IEEE Transactions on Automatic Control,* **37**, 1022–1026.

Shen, J. C., et. al., 1991, Memoryless stabilization of uncertain dynamic delay systems: Riccati equation approach, *IEEE Transactions on Automatic Control,* **36**, 638–640.

Verriest, E. I., Fan, M. K. H. and Kullstam, J., 1993, Frequency domain robust stability criteria for linear delay systems, *Proc. 32nd IEEE Conf. Decision and Control,* San Antonio, TX, December.

Wang, S. S., Chen, B. S. and Lin, T. P., 1987, Robust stability of uncertain time-delay systems, *International Journal of Control,* **46**, 963-976.

Xie, L. and de Souza, C. E., 1993, Robust stabilization and disturbance attenuation for linear uncertain systems, *IEEE European Control Conference,* Groningen, the Netherlands, 667-672.

Xu, B. and Liu, Y., 1994, An improved Razumikhin-type theorem and its applications, *IEEE Trans. Automatic Control,* **39**, 839-841.

Control of Dissynchronized Systems of Difference Equations with Delay

A. S. C. Sinha
Purdue University at Indianapolis
Indianapolis, Indiana, 46202 USA

SUMMARY

A class of nonlinear difference equations with delay under arbitrary dissynchronization and uncertainty is examined. The stability results obtained for a class of uncertain systems of nonlinear delay-difference equations are analogous to results for delay-differential equations. The dissynchronized system characterized by nonlinear delay-difference equations has not been adequately considered. We consider nonlinear delay-difference equations

$$x(n+1) = A_0 x(n) + \sum_{j=1}^{m} A_j x(n-j) + Bu(n) + f(n, x(n)), \quad n = 0, 1, \ldots \tag{1}$$

The switching instant's $x(t_n)$, $n = 0, 1, \ldots$ are denoted as $x(n)$, $n = 0, 1, 2 \ldots$ Here $x, f \in E^n$, and A_0, and $A_j(n) \in E^{n \times n}$, are real matrices and ε is a small parameter. A system is said to synchronized when all components are switched simultaneously; otherwise, the system is said to be dissynchronized, either in phase or in frequency. It is known that the appearance of even small dissynchronizations can affect significantly the system operation, rendering it either stable or unstable [1]. The basic method is to study the eigenvalue behavior of a dissynchronized linear system. We will use Liapunov-Razumikhin's type conditions for a class of nonlinear delay-difference systems. We develop a method to analyze robustness of the dissynchronized nonlinear delay-difference systems. Robustness means that the control u(n) is able to maintain system stability in the presence of the uncertainty function f(n,x(n)) for the dissynchronized delay-difference systems. It is shown that the solution is synchronized for any delay term and any uncertainty function under certain condition. In order to examine the synchronization of equation (1), we define a transformation as

$$X(n-j) = x(n-j) - \nu_j\, x(n); \quad 0 \le \nu_j < 1 \tag{2}$$

where ν_j ($j = 1, \ldots, m$) are parameters. Substituting equation (2) in (1) yields:

$$x(n+1) = [A_0 + \sum_{j=1}^{m} \nu_j A_j] x(n) + \sum_{j=1}^{m} A_j X(n-j) + Bu(n) + f(n, x(n)) \tag{3}$$

Let the nonlinear uncertainty function f(n, x(n)) satisfies

$$\| f(n, x(n)) \| \le \mu \, \|x(n)\|, \qquad \mu \ge 0 \tag{4}$$

where μ is an unknown perturbation bound obtained in Theorem 1. The vector norm and matrix norm are defined as

$$\|x\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \quad \text{and} \quad \|A\| = \left(\lambda_{\max}(A^T A) \right)^{1/2}$$

If at the switching surface $S(n) = \Omega x(n) = 0$, the equivalent $u_{eq}(n)$ is obtained by letting $S(n+1) = 0$ [2] as

$$u_{eq}(n) = -\,[(\Omega B)^{-1}(\Omega \tilde{A}_0)] x(n) \tag{5}$$

such that Ω is the solution of the matrix equation

$$P\Omega - \Omega \tilde{A}_0 = -I \tag{6}$$

Then equation (1) can be written as

$$\begin{bmatrix} x(n+1) \\ s(n+1) \end{bmatrix} = \begin{bmatrix} \left(I - B(\Omega B)^{-1}\Omega\right)\left[A_0 + \sum_{j=1}^{m} v_j A_j\right] & B(\Omega B)^{-1}C \\ 0 & C \end{bmatrix} \begin{bmatrix} x(n) \\ s(n) \end{bmatrix}$$

$$+ \begin{bmatrix} \sum_{j=1}^{m} A_j & 0 \\ \sum_{j=1}^{m} \Omega A_j & 0 \end{bmatrix} \begin{bmatrix} x(n-j) \\ s(n-j) \end{bmatrix} + \begin{bmatrix} f(n, x(n)) \\ \Omega f(n, x(n)) \end{bmatrix} \tag{7}$$

Define a state-space as $z(n) \equiv [x(n)\ s(n)]^T$. Then equation (7) can be written as

$$z(n+1) = \widetilde{A}_0 z(n) + \sum_{j=1}^{m} \widetilde{A}_j Z(n-j) + \Delta\widetilde{f}(n) \tag{8}$$

The coefficients matrices $\tilde{A}_0$, $\tilde{A}_j$ and $Z(n-j)$ are defined by equation (7). Next, we have the following theorem in presence of an uncertainty function $f(n, x(n))$:

Theorem 1. Let there exist positive-definite matrices P and Q, and constants $0 \le v_j < 1$, $j = 1, \ldots, m$ such that the following inequality holds:

$$\text{(i)}\ \alpha \equiv \lambda_{\min}(Q) - \ell(P) > 0$$

where

$$\ell(P) \equiv 2\sum_{j=1}^{m} \left\| \widetilde{A}_0^T P \widetilde{A}_j \right\| \left[v_j + \left(\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)} \right)^{\frac{1}{2}} \right]$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{m} \left\| \widetilde{A}_i^T P \widetilde{A}_j \right\| \left[v_i + \left(\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)} \right)^{\frac{1}{2}} \right] \left[v_j + \left(\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)} \right)^{\frac{1}{2}} \right]$$

In addition if the uncertainty function $f(n, x(n))$ satisfy the inequality

$$\text{(ii)}\ \|f(n, x(n))\| \le \mu\, \|x(n)\|,\ \mu \ge 0,$$

such that

$$\text{(iii)}\ 0 \le \mu < \left(\frac{\lambda_{\min}(P)}{\lambda_{\max}(P)} \right)^{\frac{1}{2}} \frac{[\lambda_{\max}(P-Q) + \alpha]^{\frac{1}{2}} - [\lambda_{\max}(P-Q)]^{\frac{1}{2}}}{[\alpha + \lambda_{\max}(\Omega^T \Omega)]^{\frac{1}{2}} [\lambda_{\max}(P)]^{\frac{1}{2}}}$$

Then system (8) is stable for any dissynchronizations such that $\| x(n)\| < \varepsilon$, provided $\|x(n_0)\| < \delta(\varepsilon)$,where $\delta(\varepsilon) = \varepsilon \Big/ \left(\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)} \right)^{\frac{1}{2}}$.

REFERENCES

[1] A. F. Kleptsyn, et al., (1984). The effect of a small dissynchronization on the stability of complex systems. *Automatika i telmekhanika,* **3**, 42-47.

[2] S.V. Drakunov, and V. I. Utkin,. (1989). On discrete-time sliding modes, *Proceedings of IFAC Conference on Nonlinear Control System Design,* Capri, Italy, 273-277.

Sufficient Conditions for the Exponential Stability of Linear Systems Subject to Norm Bounded, Bounded Rate Uncertainties

Francesco Amato

Dipartimento di Informatica e Sistemistica
Università degli Studi di Napoli "Federico II" via Claudio 21, 80125 Napoli, Italy
Email: amato@disna.dis.unina.it

1 Introduction and Problem Statement

In this paper we consider the following linear system subject to norm bounded, bounded rate uncertainties

$$\dot{x}(t) = (A + B\Delta(t)C)x(t) \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is Hurwitz, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{r\times n}$ and

$$\Delta(\cdot) \in \mathcal{D} := \left\{ \Delta(\cdot) \in \mathcal{C}^1_{m\times r}\,,\ \Delta^T(t)\Delta(t) \le I\,,\ \dot{\Delta}^T(t)\dot{\Delta}(t) \le D\,,\ t \in [0, +\infty) \right\}. \tag{2}$$

In (2) $\mathcal{C}^1_{m\times r}$ denotes the set of continuously differentiable matrix-valued functions taking values in $\mathbb{R}^{m\times r}$ and D is a given positive definite matrix.

The term "norm bounded" follows from the fact that the uncertainties satisfy the normalized bound (the scaling factors are included in the matrices B and C)

$$\Delta^T(t)\Delta(t) \le I\,; \tag{3}$$

the term "bounded rate" follows from the condition

$$\dot{\Delta}^T(t)\dot{\Delta}(t) \le D \tag{4}$$

which represents an upper bound on the rate of variation of $\Delta(\cdot)$.

Definition 1 [4]
System (1) is said to be quadratically stable if there exists a positive definite symmetric matrix P such that for all Δ with $\Delta^T\Delta \le I$

$$(A + B\Delta C)^T P + P(A + B\Delta C) < 0 \tag{5}$$

■

It is readily seen that quadratic stability guarantees uniform exponential stability of system (1) for arbitrarely time-varying uncertainties satisfying the bound (3). The quadratic stability test can be reduced to a feasibility problem involving Linear Matrix Inequalities (LMIs); such feasibility problem can be easily solved by one of the algorithms proposed in [5].

The simplicity of the quadratic stability test, rendered this approach quite popular in the control community in the last years. At the same time it is clear that such approach may be conservative when the rate of variation of the uncertainties is bounded, as in the case of this paper.

The goal of this paper is that of stating sufficient conditions for the stability of system (1) taking into account the information (4).

The previous literature considered only the case in which Δ is a *diagonal matrix* (see [6], [2] and [3]); in that case the uncertainties reduce to a vector of uncertain parameters which attains values into a given hyper-rectangle. In this paper we consider the more general *full matrix* case with the classical norm bound (3).

The stability analysis will be performed with the aid of Δ-dependent Lyapunov functions in the form

$$v(x, \Delta) = x^T P(\Delta) x \tag{6}$$

where

$$P(\Delta) = P_0 + N \Delta C \tag{7}$$

In (7) $P_0 \in \mathbb{R}^{n \times n}$ and $N \in \mathbb{R}^{n \times m}$ are left free for optimization purposes; P_0 is required to be a symmetric matrix. Note that, in some sense, the Lyapunov function (6) extends the structure of the Lur'è-Postnikov Lyapunov function to the full matrix case.

2 Main Result

The following theorem is the main result of the paper (for a proof see [1]).

Theorem 1 *System (1) is uniformly exponentially stable for all $\Delta(\cdot) \in \mathcal{D}$ if there exist*

i) A symmetric matrix $P_0 \in \mathbb{R}^{n \times n}$;

ii) A matrix $N \in \mathbb{R}^{n \times m}$;

iii) Nonnegative scalars τ_0, τ_1, τ_2

such that

$$\begin{pmatrix} P_0 - \tau_0 C^T C & \frac{1}{2}N \\ \frac{1}{2}N^T & \tau_0 I \end{pmatrix} > 0 \tag{8a}$$

$$\begin{pmatrix} A^T P_0 + P_0 A + \tau_1 C^T C + \tau_2 C^T D C & A^T N + P_0 B & \frac{1}{2}N \\ N^T A + B^T P_0 & B^T N + N^T B - \tau_1 I & O \\ \frac{1}{2}N^T & O & -\tau_2 I \end{pmatrix} < 0 \tag{8b}$$

■

By virtue of Theorem 1, we have reduced the problem of testing exponential stability of system (1) subject to norm bounded, bounded rate uncertainties to a LMIs based feasibility problem.

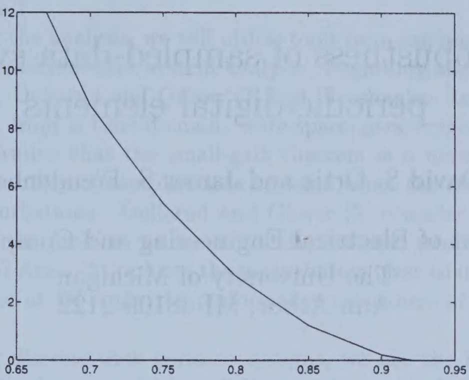


Figure 1: x-axis: ρ, y-axis: $\overline{\rho}_{max}$

3 Example

Consider the uncertain system (1)–(4) with

$$A = \begin{pmatrix} 0 & 1 \\ -5 & -2 \end{pmatrix} \qquad B = \rho \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad D = \overline{\rho} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} . \tag{9}$$

This uncertain system is not quadratically stable for $\rho > 2/3$; therefore, for any given $\rho > 2/3$, it makes sense to use Theorem 1 to find an estimate of the maximum $\overline{\rho}$ such that system (9) is exponentially stable. The results are depicted in Figure 1.

References

[1] Amato F., "Sufficient conditions for the exponential stability of linear systems subject to norm bounded, bounded rate uncertainties," *Internal Report*, Università di Napoli, July 1997.

[2] Amato F., M. Corless, M. Mattei, R. Setola, "Robust stability of linear systems depending on Uncertain, slowly-varying parameters," *In Proc. 33rd Allerton Conference*, Monticello (IL), Oct. 1995.

[3] Amato F., M. Corless, M. Mattei, R. Setola, "A robust stability margin in the presence of time varying, bounded rate gains," *Int. J. Robust and Nonlinear Control*, Vol 7, pp. 127–143, Feb. 97.

[4] Barmish B. R., " Stabilization of uncertain systems via linear control," *IEEE Trans. Aut. Contr.*, vol. AC-28, pp. 848–850, 1983.

[5] Boyd S., L. El Ghaoui, E. Feron, V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Press, 1994.

[6] Jonsson U. and A. Rantzer. Systems with uncertain parameters - Time-variations with bounded derivatives. *In Proc. 33rd Conference on Decision and Control*, Orlando (FL), Dec. 1994.

Stability robustness of sampled-data systems with periodic-digital elements

David S. Ortiz and James S. Freudenberg

Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109-2122

Abstract

A framework for the frequency-domain analysis of sampled-data systems is presented and within it, necessary and sufficient conditions for stability robustness to multiplicative perturbations for sampled-data systems with periodic-digital elements are presented. Multiplicative perturbations to the continuous-time dynamics are considered. Control of air/fuel ratio in a four-cylinder engine is presented as an example to illustrate the results; we show that the system is robustly stable to an unknown delay of 0.0076 seconds.

1 Introduction and motivation

Consider the sampled-data system with periodic-digital controller depicted in Figure 1. It is a model of an idle speed control loop for a fuel injected engine and has a unique structure [10]. Although the four cylinders are independent dynamic systems, their dynamics are sufficiently fast with respect to the manifold dynamics that they may be considered to occur discretely; furthermore, since each cylinder fires in sequence, the engine, is inherently discrete and periodic. The transport dynamics, however, are continuous. Hence we have a system which inherently exhibits the characteristics of a sampled-data system such that the discrete components are periodic.

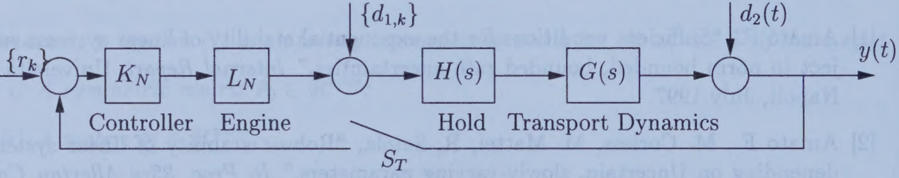


Figure 1: Sampled-data system with periodic-discrete plant

The control problem is to design a controller which regulates the air-fuel ratio in each cylinder. The problem may be solved through application of periodic-control synthesis and analysis techniques [4, 12]. In the continuous transport dynamics, however, there is an analog delay. Let $G(s) = G_0(s)e^{-s\tau}$, the value of the delay, τ, is poorly known, and variation around the nominal value may cause performance and robustness problems.

Modeling the delay as a multiplicative uncertainty, we will study the stability robustness of the system in Figure 1.

In order to perform the analysis, we will utilize tools from sampled-data system analysis and from periodic discrete-time system analysis. Regarding stability robustness for sampled-data systems, Dullerud and Glover [3] and Sivashankar and Khargonekar [15] approach the problem from a time-domain, state-space perspective. Sivashankar and Khargonekar [15] determine that the small-gain theorem is a necessary and sufficient condition for robust stability of sampled-data systems when the class of perturbations includes periodic perturbations. Dullerud and Glover [3] consider time-invariant perturbations and reduce the problem to an infinite-dimensional structured singular value problem. Hagiwara and Araki [11] extend the perturbation class to include unstable perturbations. Grasselli, *et al.* [9] study the performance robustness of linear discrete-time periodic systems.

For the analysis of discrete-time periodic systems, we use the techniques of Ortiz, *et al.* [13], especially the frequency-lifting of discrete-time periodic systems [7, 8]. The frequency-domain techniques for sampled-data systems from Freudenberg, *et al.* [5, 6] are the basis of our analysis approach. Together, they form a method of frequency-domain analysis for the system in Figure 1. The results are extensions of the stability robustness results for sampled-data systems as presented in Braslavksy [1]. We consider only multiplicative perturbations to either the transport dynamics or the periodic-digital plant.

The paper is organized as follows. Section 2 defines the notation. Section 3 presents the necessary and sufficient conditions for robust stability to multiplicative perturbations to the continuous part of the system. While we extend a necessary and sufficient condition for robust stability based upon the structured singular value from Dullerud and Glover [3], but our main results are not both necessary and sufficient due to the complexities introduced by the periodic-digital components. Also note that the necessary condition which we develop depends upon the weighting function and the fundamental complementary sensitivity function for periodic-digital sampled-data systems. It is shown in Ortiz, *et al.* [13] that the fundamental complementary sensitivity function has integral constraints due to time delays. Section 4 presents the application of the results in Section 3 to the problem in Grizzle, *et al.* [10].

2 Notation and preliminary results

In this section, we present the analysis tools that we require to develop the results in later sections. For a complete development of a frequency-based analysis technique for sampled-data systems with periodic-digital elements, see Ortiz, *et al.* [13]. We begin by redrawing the system in Figure 1 in a more general form. Figure 2 is the system upon which we will base our analysis; it differs from Figure 1 in that we lump the periodic controller and periodic plant into one block and represent the continuous dynamics by $P(s)$; $P(s) = P_0(s)e^{-s\tau}$, $\tau > 0$, where $P_0(s)$ is rational. In addition, we have incorporated multiplicative uncertainty of the plant into the model.

For any function, $P(s)$ of the form above, let

$$P_{k,l}(s) := P\left(s + j(k\omega_s - l\frac{\omega_s}{N})\right), \tag{1}$$

where k is an integer, $l \in \{0, 1, \ldots, N-1\}$, ω_s is the sampling frequency, and N is the period of the underlying periodic digital system. Define the $N \times N$ diagonal matrix

$$\underline{P}_k(s) := \operatorname{diag}\{P_{k,0}(s), P_{k,1}(s), \ldots, P_{k,N-1}(s)\}. \tag{2}$$

For continuous signals, $Y(s)$, define

$$\underline{Y}_k(s) := \begin{bmatrix} Y_{k,0}(s) & Y_{k,1}(s) & \cdots & Y_{k,N-1}(s) \end{bmatrix}^T. \tag{3}$$

For discrete signals, $U_d(z)$ define

$$\underline{U}_d(z) := \begin{bmatrix} U_d(z) & U_d(ze^{-j\frac{\omega_s}{N}T}) & \cdots & U_d(ze^{-j(N-1)\frac{\omega_s}{N}T}) \end{bmatrix}^T. \tag{4}$$

The representation of the analog transfer function in terms of (2) will allow us to derive a new expression for the closed loop response of sampled-data systems with periodic-digital elements; from Ortiz, *et al.* [13], recall that if C_N represents a discrete-time system of period N, then $C_{freq}(z)$ is its frequency-lifted representation [7]. The following relationship between elements of a frequency-lifted periodic-digital system is used extensively in the sequel.

Lemma 1 ([13]). *Let $C^{k,l}_{freq}(z)$ be the $(k+1, l+1)$ element of the frequency-lifted matrix $C_{freq}(z)$, $k, l \in \{0, \ldots, N-1\}$ and $\omega_p := 2\pi/N$. Then*

$$C^{l,l}_{freq}(z) = C^{0,0}_{freq}(ze^{-jl\omega_p}) \tag{5}$$

$$C^{0,l}_{freq}(z) = C^{N-l,0}_{freq}(ze^{-jl\omega_p}) \tag{6}$$

$$C^{l,0}_{freq}(z) = C^{0,N-l}_{freq}(ze^{-jl\omega_p}) \tag{7}$$

Lemma 1 relates the elements in the first row of a frequency lifted representation to the elements of the remaining $N-1$ rows. Let C_{freq} and S_{freq} be two frequency-lifted representations of N-periodic systems. Let $(CS)_{freq}$ represent the product $C_{freq}S_{freq}$ and $(CS)^{k,l}_{freq}$ represent the $(k+1, l+1)$ element of $(CS)_{freq}$, where $k, l \in \{0, 1, \ldots, N-1\}$.

If $T > 0$ is the sampling period, then the sampling frequency, ω_s, is $2\pi/T$; if a positive integer N is the period of a discrete-time system, then its frequency of oscillation, ω_p, is $2\pi/N$. Note that $\omega_p N = \omega_s T$. Let $\mathbb{D}$ be the open unit disc in the complex plane and $\mathbb{D}^c$ its complement. Let $\Omega_N := [0, \omega_s/2)$.

3 Stability robustness to LTI multiplicative uncertainty

The problem as stated in Section 1 is to determine the values of the delay for which the system will be robustly stable. We consider the more common problem of stability robustness to multiplicative perturbations in the continuous plant; in Section 4 we model the delay as a multiplicative perturbation. The perturbation is of the form,

$$\tilde{P}(s) = (1 + W(s)\Delta(s))\,P(s), \tag{8}$$

where $P(s)$ is the modeled plant, and $\Delta(s)$ is a perturbation given by a stable rational function such that $\|\Delta\|_\infty < 1$. A Δ which satisfies these assumptions is called an admissible perturbation. The weighting function, $W(s)$, is a stable, minimum phase

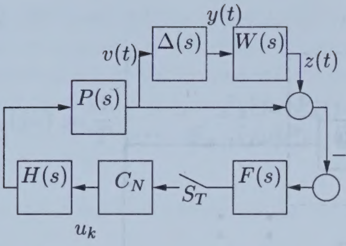


Figure 2: Block diagram of multiplicative uncertainty model for periodic control

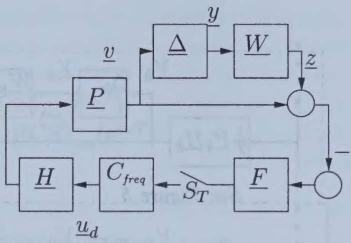


Figure 3: Block diagram of multiplicative uncertainty model for periodic control written as a MIMO lifted system

transfer function such that the product $F(s)W(s)P(s)$ is proper. These assumptions are consistent with those from Braslavsky [1].

Assume for the nominal plant, $P(s)$, and controller, C_N, that the system is stable. In the sequel, we will develop necessary and sufficient conditions for the system to remain stable under admissible perturbations.

Recall $C_{freq}(z)$ is the frequency-lifted representation of C_N and $S_{freq}(z)$ is a frequency-lifted representation of the discrete sensitivity function, $(I + C_{freq}(z)P_{freq}(z))^{-1}$, where $P_{freq}(z)$ is the frequency-lifted representation of the discretized plant. Then we may redraw Figure 2 as a MIMO system. Figure 3 is a diagram of the MIMO representation of the sampled-data system with periodic digital controller. Many researchers have studied periodic systems from their corresponding MIMO time-invariant representations, generalizing results accordingly [7, 8, 9]. We will first generalize MIMO results to the corresponding frequency-lifted system and then use the structural properties of lifted systems, as illustrated in Lemma 1, to simplify and interpret the results.

In order to gain an understanding for the coupling of frequencies in sampled-data systems with periodic digital elements, consider Figure 4. Figure 4 is a diagram of the frequency interactions of a sampled-data system and is a generalization of Figure 6.4 from Braslavsky [1], who considered only shift-invariant digital controllers. Note that for each aliased frequency due to the sampled-data feedback, there exist N independent feedback paths separated in frequency by $\omega_p/T = \omega_s/N$.

Figures 4 and 5 illustrate the dual-scale interaction of the frequencies for periodic control: the first frequency scale is ω_s, the sampling frequency, and is a consequence of sampled-data control; the second frequency scale is $\frac{\omega_s}{N}$, which is the frequency separation between the harmonics of the discrete-time periodic system.

Using the development of Chapter 5 of Braslavsky [1], we write the closed loop as an infinite dimensional matrix equation.[1] The representation will allow us to generalize the results from Dullerud and Glover [3] to the case of periodic-digital control. Define $\underline{Z}'_k(j\omega) := \underline{F}_k(j\omega)\underline{Z}_k(j\omega)$ and

$$z_\omega(j\omega) := \left[\cdots \quad \underline{Z}'_1(j\omega)^T \quad \underline{Z}'_0(j\omega)^T \quad \underline{Z}'_{-1}(j\omega)^T \quad \cdots\right]^T, \tag{9}$$

---

[1] Braslavsky [1] refers to the infinite dimensional matrix representation of the system as the *frequency lifting* of the system. We have already reserved the term *frequency lifting* to describe a matrix representation of a periodic-digital system.

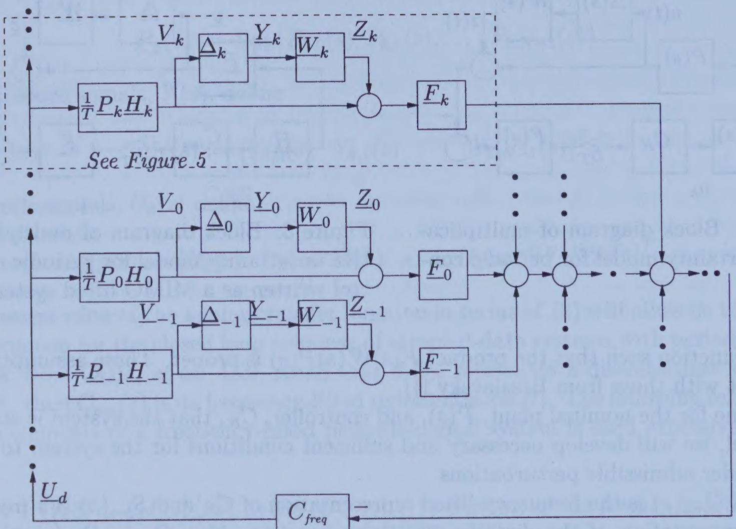

$\underline{V}_k$
$\underline{\Delta}_k$
$\underline{Y}_k$
$\underline{W}_k$
$\underline{Z}_k$
$\frac{1}{T}\underline{P}_k\underline{H}_k$
$\underline{F}_k$
See Figure 5
$\underline{V}_0$
$\underline{\Delta}_0$
$\underline{Y}_0$
$\underline{W}_0$
$\underline{Z}_0$
$\frac{1}{T}\underline{P}_0\underline{H}_0$
$\underline{F}_0$
$\underline{V}_{-1}$
$\underline{\Delta}_{-1}$
$\underline{Y}_{-1}$
$\underline{W}_{-1}$
$\underline{Z}_{-1}$
$\frac{1}{T}\underline{P}_{-1}\underline{H}_{-1}$
$\underline{F}_{-1}$
$\underline{U}_d$
C_{freq}
$-$

Figure 4: Diagram of frequency interactions for periodic control

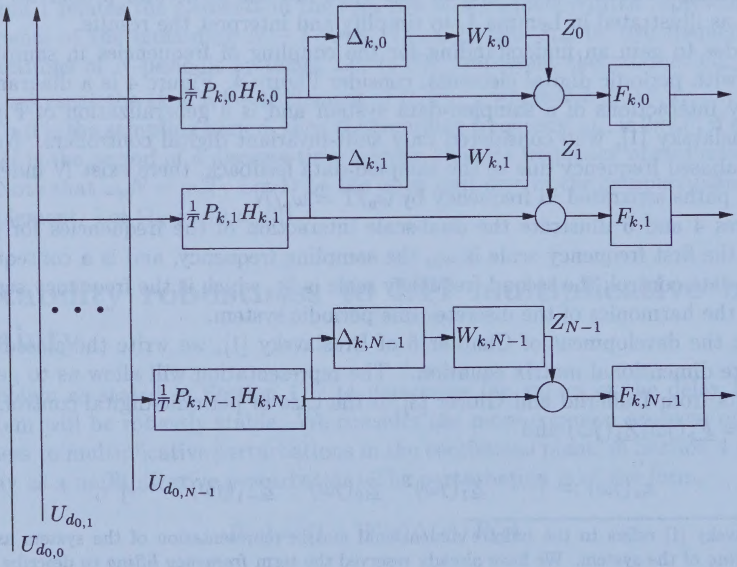

$\Delta_{k,0}$
$W_{k,0}$
Z_0
$\frac{1}{T}P_{k,0}H_{k,0}$
$F_{k,0}$
$\Delta_{k,1}$
$W_{k,1}$
Z_1
$\frac{1}{T}P_{k,1}H_{k,1}$
$F_{k,1}$
$\Delta_{k,N-1}$
$W_{k,N-1}$
Z_{N-1}
$\frac{1}{T}P_{k,N-1}H_{k,N-1}$
$F_{k,N-1}$
$U_{d_{0,N-1}}$
$U_{d_{0,1}}$
$U_{d_{0,0}}$

Figure 5: Harmonic structure of forward loop k

and

$$T_\omega(j\omega) := \frac{1}{T}\left[\begin{array}{cc} & \vdots \\ \cdots & \underline{F}_k(j\omega)\underline{P}_k(j\omega)\underline{H}_k(j\omega)(CS)_{\mathit{freq}}(e^{j\omega T}) \\ \cdots & \underline{F}_{k-1}(j\omega)\underline{P}_{k-1}(j\omega)\underline{H}_{k-1}(j\omega)(CS)_{\mathit{freq}}(e^{j\omega T}) \\ & \vdots \end{array}\right.$$

$$\left.\begin{array}{cc} \vdots & \\ \underline{F}_k(j\omega)\underline{P}_k(j\omega)\underline{H}_k(j\omega)(CS)_{\mathit{freq}}(e^{j\omega T}) & \cdots \\ \underline{F}_{k-1}(j\omega)\underline{P}_{k-1}(j\omega)\underline{H}_{k-1}(j\omega)(CS)_{\mathit{freq}}(e^{j\omega T}) & \cdots \\ \vdots & \end{array}\right], \qquad (10)$$

$$W_\omega(j\omega) := \operatorname{diag}\{\cdots, \underline{W}_k(j\omega), \underline{W}_{k-1}(j\omega), \cdots\}, \qquad (11)$$

$$\Delta_\omega(j\omega) := \operatorname{diag}\{\cdots, \underline{\Delta}_k(j\omega), \underline{\Delta}_{k-1}(j\omega), \cdots\}. \qquad (12)$$

Lemma 2. *The closed-loop response of the system in Figure 2 is characterized by*

$$(I + \Delta_\omega(j\omega)W_\omega(j\omega)T_\omega(j\omega))\, z_\omega(j\omega) = 0. \qquad (13)$$

Remark 1. The infinite dimensional matrix equation, (13), has a feedback interpretation as a standard perturbation model as in Figure 6.

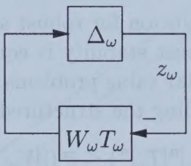


Figure 6: Infinite dimensional perturbation model

Note that Figure 6 and (13) satisfy the assumptions of the structured singular value problem as developed in Packard and Doyle [14] and its application to the sampled-data stability robustness analysis of Dullerud and Glover [3]. The following proposition will allow us to perform a robust stability analysis of the system in Figure 2.

We state the necessary and sufficient condition for stability robustness in terms of the structured singular value of a system which is truncated in frequency. Define truncated versions of T_ω, W_ω and Δ_ω (omitting arguments):

$$[T_\omega]^n = \frac{1}{T}\begin{bmatrix} \underline{F}_n\underline{P}_n\underline{H}_n(CS)_{\mathit{freq}} & \cdots & \underline{F}_n\underline{P}_n\underline{H}_n(CS)_{\mathit{freq}} \\ \vdots & & \vdots \\ \underline{F}_{-n}\underline{P}_{-n}\underline{H}_{-n}(CS)_{\mathit{freq}} & \cdots & \underline{F}_{-n}\underline{P}_{-n}\underline{H}_{-n}(CS)_{\mathit{freq}} \end{bmatrix}, \qquad (14)$$

$$[W_\omega]^n = \operatorname{diag}\{\underline{W}_n, \cdots, \underline{W}_{-n}\}, \qquad (15)$$

$$[\Delta_\omega]^n = \operatorname{diag}\{\underline{\Delta}_n, \cdots, \underline{\Delta}_{-n}\}. \qquad (16)$$

Proposition 1. *For all admissible Δ, the system of Figure 2 is internally stable if and only if, for each integer $n > 0$,*

$$\max_{\omega\in\Omega_N} \mu_{\Delta^n}\left([W_\omega(j\omega)]^n[T_\omega(j\omega)]^n\right) \le 1. \qquad (17)$$

We now derive necessary and sufficient conditions for stability robustness of the system in Figure 2; these conditions will lend insight into the structure of the system and are easier to evaluate than the structured singular value problems in Proposition 1. The derivation is simplified because the generically unstructured perturbation in Figure 2, Δ, is diagonal in the form Δ_ω. We first present a necessary condition for internal stability of the system in Figure 2.

Theorem 1. *A necessary condition for robust stability of the system in Figure 2 is*

$$\|T_{fun}(j\omega)W(j\omega)\|_\infty \leq 1, \; \forall \omega \in \Omega_N \tag{18}$$

where

$$T_{fun}(j\omega) := \frac{1}{T} F(j\omega)P(j\omega)H(j\omega)(CS)^{0,0}_{\mathit{freq}}(e^{j\omega T}). \tag{19}$$

Remark 2. Theorem 1 shows that the fundamental component of the response characterizes a necessary condition for robust stability of the closed-loop. The condition is similar to that of Braslavsky [1].

Remark 3. The fundamental complementary sensitivity function characterizes the fundamental component of the steady-state response to a sinusoidal command input. Also, as shown in Ortiz, *et al.* [13], T_{fun} satisfies an integral constraint dependent upon time delays in the continuous plant. Theorem 1 shows that results predicting peaks in T_{fun} have direct applications in stability robustness analysis.

We now present a sufficient condition for robust stability of the system in Figure 2. Recall from Proposition 1 that robust stability is equivalent to a gain condition on an infinite number of structured singular value problems. From Packard and Doyle [14], we have the following inequality regarding the structured singular value:

$$\mu_{\Delta^n}\left([W_\omega]^n[T_\omega]^n\right) \leq \overline{\sigma}\left([W_\omega]^n[T_\omega]^n\right). \tag{20}$$

We will derive an upper bound of the truncated structured singular value of $[W_\omega]^n[T_\omega]^n$. Note the structure of $[W_\omega]^n[T_\omega]^n$:

$$[W_\omega]^n[T_\omega]^n = \frac{1}{T}\begin{bmatrix} \underline{W}_n\underline{F}_n\underline{P}_n\underline{H}_n(CS)_{\mathit{freq}} & \cdots & \underline{W}_n\underline{F}_n\underline{P}_n\underline{H}_n(CS)_{\mathit{freq}} \\ \vdots & & \vdots \\ \underline{W}_{-n}\underline{F}_{-n}\underline{P}_{-n}\underline{H}_{-n}(CS)_{\mathit{freq}} & \cdots & \underline{W}_{-n}\underline{F}_{-n}\underline{P}_{-n}\underline{H}_{-n}(CS)_{\mathit{freq}} \end{bmatrix}. \tag{21}$$

Theorem 2. *Let Δ be an admissible perturbation. Then a sufficient condition for robust stability of the system in Figure 2 is that, for all $\omega \in \Omega_N$ and $n > 0$,*

$$\frac{1}{T}\sum_{l=1}^{N-1} \max_m \left\{ \sum_{k=-n}^{n} \left| F_{k,m}(j\omega)P_{k,m}(j\omega)H_{k,m}(j\omega)(CS)^{m,(m+l)|N}_{\mathit{freq}}(e^{j\omega T})W_{k,m}(j\omega) \right| \right\} \leq 1, \tag{22}$$

where $(m+l)|N$ denotes addition modulo N.

Remark 4. The sufficient condition, (22), is a complicated expression for the upper bound of a maximum singular value of a matrix. We may evaluate the expression in practice to guarantee the stability robustness of a system (see Section 4.) We have reduced the infinite-dimensional structured singular value problem to a summation of a finite-dimensional problem.

4 Example

We return to the fuel-injected engine idle speed control loop — Figure 1. Recall that this control problem has been considered and a controller has been designed using LQG/LTR methods [10]. Consider a multiplicative model for the perturbation of the delay in the sensor dynamics:

$$G(s) = G_0(s)e^{-\kappa T}e^{-s\epsilon T} = (1 + W(s)\Delta(s))G_0(s)e^{-\kappa T} \tag{23}$$

where $0 < \epsilon < 1$. The delay has a nominal value, κ, of 7 engine cycles or 0.14 seconds which is absorbed into the sensor dynamics. Here, we consider perturbations to the nominal value of the delay; we approximate the frequency response of the perturbations by the following weighting function [2]:

$$W(s) = \frac{0.0147s}{0.007s + 1} \tag{24}$$

Figure 7 is a plot of (24) and the delay parameter ϵ. We see that (24) approximates a delay of 0.38 sample periods or 0.0076 seconds. In order to define the frequency-lifted sensitivity function, lump the cascade of the periodic plant and controller into a single periodic-discrete system and let the anti-aliasing filter of Figure 2 be unity.

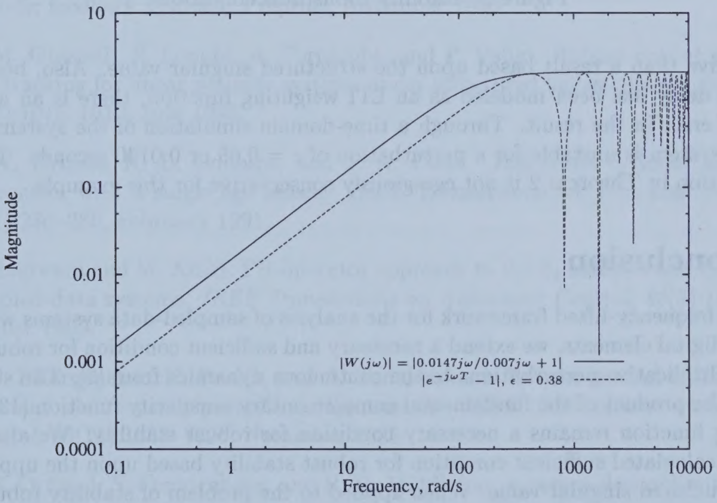


Figure 7: Weighting function

Apply the results of Section 3 to the system in Figure 1. Figure 8 is a plot of the results of Theorems 1 and 2. The system in robustly stable because the sufficient condition, (22), has a maximum value less than unity. As ω approaches the sampling frequency, the difference between conditions (18) and (22) becomes small. Robust stability conditions based upon the structured singular value such as that given in Proposition 1 are conservative because of the assumption of unstructured uncertainty; in fact, it is structured because it is a time delay. Because the sufficient condition for robust stability, (22), is based upon a bound on the structured singular value, we expect it to be more

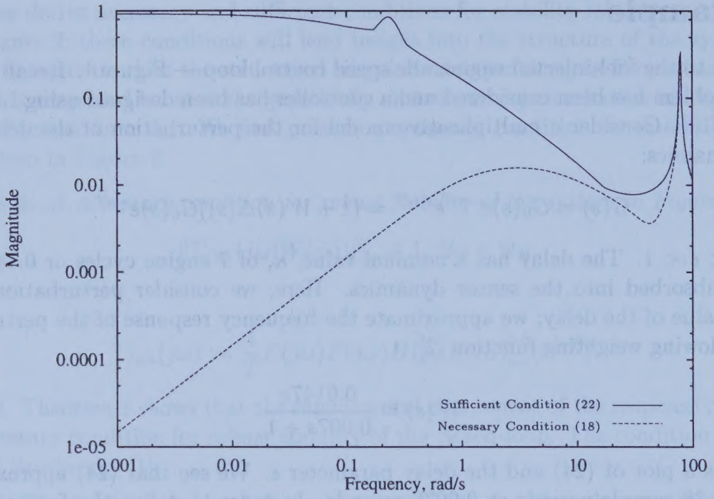


Figure 8: Stability robustness conditions

conservative than a result based upon the structured singular value. Also, because the unknown delay has been modeled as an LTI weighting function, there is an additional source of error in the result. Through a time-domain simulation of the system, we find that the system is unstable for a perturbation of $\epsilon = 0.65$ or 0.0130 seconds. Therefore, the condition in Theorem 2 is not egregiously conservative for this example.

5 Conclusion

Within a frequency-lifted framework for the analysis of sampled-data systems with linear periodic digital elements, we extend a necessary and sufficient condition for robust stability to multiplicative perturbations to the continuous dynamics from [3]. The supremum norm of the product of the fundamental complementary sensitivity function [13] and the weighting function remains a necessary condition for robust stability. We also present an easily calculated sufficient condition for robust stability based upon the upper bound of the structured singular value. When applied to the problem of stability robustness of an air/fuel ratio control loop in a fuel injected engine, the sufficient condition predicts stability robustness to all delays up to 0.38 sample periods, or 0.0076 seconds, in duration, while time-simulation predicts stability robustness to all delays up to 0.65 sample periods, or 0.0130 seconds.

References

[1] J. H. Braslavsky. *Frequency domain analysis of sampled-data control systems.* PhD thesis, University of Newcastle, New South Wales 2308, Australia, October 1995.

[2] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum. *Feedback Control Theory.* Macmillan, New York, 1992.

[3] G. Dullerud and K. Glover. Robust stabilization of sampled-data systems to structured LTI perturbations. *IEEE Transactions on Automatic Control*, 38(10):1497–1508, October 1993.

[4] B. A. Francis and T. T. Georgiou. Stability theory for linear time-invariant plants with periodic digital controllers. *IEEE Transactions on Automatic Control*, 33(9):820–832, September 1988.

[5] J. S. Freudenberg, R. H. Middleton, and J. H. Braslavsky. Robustness of zero-shifting via generalized sampled-data hold functions. Technical Report EE9431, University of Newcastle, NSW 2308 Australia, August 1994.

[6] J. S. Freudenberg, R. H. Middleton, and J. H. Braslavsky. Inherent design limitations for linear sampled-data feedback systems. *International Journal of Control*, 61(6):1387–1421, 1995.

[7] G. C. Goodwin and A. Feuer. Linear periodic control: a frequency domain viewpoint. *Systems and Control Letters*, 19(5):379–390, 1992.

[8] G. C. Goodwin and G. I. Gomez. Frequency domain design trade-offs for linear periodic feedback systems. Preprint, November 1995.

[9] O. M. Grasselli, S. Longhi, A. Tornambè, and P. Valigi. Robust output regulation and tracking for linear periodic systems under structured uncertainties. *Automatica*, 32(7):1015–1019, July 1996.

[10] J. W. Grizzle, K. L. Dobbins, and J. A. Cook. Individual cylinder air-fuel ratio control with a single ego sensor. *IEEE Transactions on Vehicular Technology*, 40(1):280–286, February 1991.

[11] T. Hagiwara and M. Araki. FR-operator approach to the h_2 analysis and synthesis of sampled-data systems. *IEEE Transactions on Automatic Control*, 40(8):1411–1421, August 1995.

[12] P. P. Khargonekar, K. Poolla, and A. Tannenbaum. Robust control of linear time-invariant plants using periodic compensation. *IEEE Transactions on Automatic Control*, 30(11):1088–1096, November 1985.

[13] D. S. Ortiz, J. S. Freudenberg, and R. H. Middleton. Inherent design limitations for feedback systems with periodic digital controllers. Technical report, University of Michigan, Department of Electrical Engineering and Computer Science, Ann Arbor, MI 48109-2122, 1997.

[14] A. Packard and J. Doyle. The complex structured singular value. *Automatica*, 29(1):71–109, 1993.

[15] N. Sivashankar and P. P. Khargonekar. Robust stability and performance analysis of sampled-data systems. *IEEE Transactions on Automatic Control*, 38(1):58–69, January 1993.

A New Approach to the Synthesis of Parameter Dependent Lyapunov Functions for the Estimation of the Stability Margin in the Presence of Slowly-Varying Parameters

F. Amato and G. Celentano

Dipartimento di Informatica e Sistemistica Università degli Studi di Napoli "Federico II"
via Claudio 21, 80125 Napoli, Italy
Email: amato@disna.dis.unina.it

Abstract

In this paper we consider a MIMO asymptotically stable linear plant. For such a system the classical concepts of *quadratic stability margin* and *multivariable gain margin* can be defined. These margins have the following interpretation: consider the closed loop system composed of the plant and some real parameters, one inserted in each channel of the loop; then any time-varying (time-invariant) parameters whose amplitudes are smaller than the quadratic stability (multivariable gain) margin result in a stable closed loop system. For time-varying parameters whose magnitudes are between these two stability measures, stability may depend on the *rate of variation* of the parameters. Therefore it makes sense to consider the stability margin given by the maximal allowable rate of variation of the parameters which guarantees stability of the closed loop system. As shown in previous papers, an estimate of this margin can be obtained with the aid of parameter dependent Lyapunov functions. In this paper we propose a new approach to the search of parameter dependent Lyapunov functions which leads to a *much less conservative* algorithm to estimate such margin at the price of a greater computational burden; this limits the application of the methodology to systems containing up to three parameters. In this context, an example will show that the proposed methodology introduces a strong improvement with respect to existing techniques.

1 Introduction

In this paper we consider the closed loop system composed of the nominal, asymptotically stable MIMO plant described by the triple $(\boldsymbol{A}_0, \boldsymbol{B}, \boldsymbol{C})$ subject to some feedback loops. The gain of each feedback loop is independent and is given by an uncertain real parameter.

Referring to this setup, there are two classical robustness measures. An uncertain system is said to be quadratically stable when there exists an uncertainty independent quadratic Lyapunov function whose derivative along the solutions of the system is negative definite for all allowable uncertainties; in this framework the *quadratic stability margin* (see [12]) is defined as the smallest amplitude of the parameters such that the overall closed loop system is not quadratically stable. Quadratic stability guarantees uniform asymptotic stability in the presence of *time-varying* parameters (see [10]). The second robustness measure we consider is the *multivariable gain margin*, defined as the smallest amplitude of *time-invariant* parameters which results in an unstable closed loop system (see [16]).

For time-varying parameters whose magnitudes are between these two stability measures (in the sequel slowly-varying parameters), stability may depend on the rate of variation of the parameters (see [15] and [4]); therefore it makes sense to consider another stability margin defined as the *maximal allowable rate of variation* of the parameters which results in a stable closed loop system (see [5]). As shown in previous works [6], [5], an estimate of such a margin can be obtained with the aid of *parameter dependent* quadratic Lyapunov functions in the form $v(x, p) = x^T P(p)x$, where p denotes the vector of parameters.

The first explicit use of parameter dependent Lyapunov functions, to evaluate the stability of systems subject to bounded rate parameters, has been done in [7]. In that paper a system subject to only one parameter was considered and the parameter dependent Lyapunov function was optimized over the class of Lur'è-Postnikov functions (a strict subclass of the affinely dependent Lyapunov functions). The extension of the results contained in [7] to the multiparameter case has been done in [6]; in both cases the problem of finding an optimal $P(\cdot)$ was reduced to a feasibility problem involving Linear Matrix Inequalities (LMIs) [12] using the $\mathcal{S}$-procedure approach which allows to transform (without introducing conservatism) the terms of the Lyapunov derivative depending on p^2 into affine terms. The more general class of affine parameter dependent Lyapunov functions has been first considered in the Doctoral Thesis [1]. Clearly the approach in [1], optimizing over a larger class of parameter dependent Lyapunov functions, revealed to be less conservative; however, since in that case the Lyapunov derivative turned out to be quadratic in the parameters (without the chance of using $\mathcal{S}$-procedure based techniques), to recover an LMIs based optimization problem the use of politopic covering techniques (see [8]), which allows to transform at the price of some conservatism a nonlinear function into a multiaffine one, was necessary; therefore the computational complexity increased. All the results of the previous cited works, plus some further insights are contained in the recent paper [5].

Independently from the above papers, in [18] and [17] the authors obtained similar results using affine parameter dependent Lyapunov functions. The difference between the approach in [5] and [18], [17] consists of the different way the nonlinear dependence on the parameters in the Lyapunov derivative is treated; some examples seem to show that the polytopic covering approach used in [5] leads to less conservative results.

Anyway, in all the above papers, systems depending affinely on the parameters were considered and similarly the structure of $P(\cdot)$ was fixed to be affine in the parameters; the coefficients of $P(\cdot)$ were left as degree of freedom for optimization purposes in order to find, if existing, a parameter dependent Lyapunov function which allows to prove the exponential stability of the system.

In this paper we present a completely different approach for the search of an optimal parameter dependent Lyapunov function, which leads to a *much less conservative* algorithm for estimating the stability margins in presence of bounded rate parameters; our approach will also show that the common idea of using parameter dependent Lyapunov functions, whose structure mimics the affine parameter dependence of the system matrix, may lead to overly conservative results. This claim will be clearly illustrated by an example provided in the final part of the paper; this example will concern a simple second order-two parameter uncertain system which will serve to clarify the application of our methodology and as benchmark to compare the proposed technique with previous ones.

Our approach has not to be considered a panacea for general robustness problems involving bounded rate parameters, because, as we will show later in the paper, the computational complexity of the proposed optimization algorithm is greater than those ones of previous papers; for systems depending on more than three parameters the techniques provided previously can be still useful.

For the sake of brevity, the proofs of Theorems are omitted; the interested reader is referred to [3].

2 Problem Statement and Preliminary Result

In this paper we consider the feedback system described by the state space equations

$$\begin{aligned} \dot{x} &= A_0 x + Bu && \text{(1a)} \\ y &= Cx && \text{(1b)} \\ u &= \Pi(p) y\,. && \text{(1c)} \end{aligned}$$

where $A_0 \in \mathbb{R}^{n \times n}$ is Hurwitz (that is, all eigenvalues of A_0 have negative real parts), $B = \begin{pmatrix} b_1 & \dots & b_q \end{pmatrix} \in \mathbb{R}^{n \times q}$, $C = \begin{pmatrix} c_1 & \dots & c_q \end{pmatrix}^T \in \mathbb{R}^{q \times n}$ and $\Pi(p) = \operatorname{diag}(p) \in \mathbb{R}^{q \times q}$ with $p := \begin{pmatrix} p_1 & \dots & p_q \end{pmatrix}^T$ being a vector of uncertain parameters.

This uncertain system can be described concisely as

$$\dot{x} = A(p)x \tag{2}$$

where

$$A(p) := A_0 + B\Pi(p)C = A_0 + \sum_{i=1}^{q} p_i b_i c_i^T \tag{3}$$

We can associate two different robustness measures with system (1):

$$\begin{aligned} \rho_Q &:= \inf\{\rho > 0 \,:\, \text{system (1) is not quadratically stable for } \|p\|_\infty \le \rho\} && (4) \\ \rho_G &:= \inf\{\|p\|_\infty \,:\, A(p) \text{ is not Hurwitz}\} && (5) \end{aligned}$$

where the infinity norm of a vector $v = \begin{pmatrix} v_1 & \dots & v_q \end{pmatrix}^T$ is given by

$$\|v\|_\infty := \max\{|v_1|, \dots, |v_q|\}.$$

The statement, "system (1) is quadratically stable for $\|p\|_\infty \le \rho$", means that there exists a quadratic Lyapunov function $V(x) = x^T P x$ whose derivative along solutions of system (1) is negative definite for all p satisfying $\|p\|_\infty \le \rho$. These two robustness measures are known in the literature as the *quadratic stability margin* [12] and the *multivariable gain margin* [16].

The computation of the quadratic stability margin can be reduced to a *generalized eigenvalue problem* (see [12], p. 66), while the computation of the multivariable gain margin can be performed according to the algorithm provided in [16].

The quadratic stability margin has a deeper meaning in the context of robust stability in the presence of *time-varying* perturbations. Indeed, system (1) is globally uniformly exponentially stable for all Lebesgue measurable, real, time-varying, parameters $p(\cdot)$, which for some $\overline{p}$, satisfy $\|p(t)\|_\infty \leq \overline{p} < \rho_Q$ for all $t \geq 0$.

Clearly, $\rho_Q \leq \rho_G$. In general this inequality is strict as many examples illustrate (see for example [14]).

In this paper, we consider the situation in which the vector p is time-varying and satisfies

$$\|p(t)\|_\infty \leq \overline{p} \quad \text{for all} \quad t \geq 0 \tag{6}$$

for some known bound $\overline{p}$. Note that the time-varying parameter $p(t)$ is constrained to belong for all t to the hypercube

$$\mathcal{S}_{\overline{p}} := [-\overline{p}, \overline{p}] \times \cdots \times [-\overline{p}, \overline{p}] = [-\overline{p}, \overline{p}]^q \subset \mathbb{R}^q$$

We denote the vertices of $\mathcal{S}_{\overline{p}}$ by $p_{(j)}$, $j = 1, \ldots, 2^q$.

If $\rho_Q = \rho_G$, then robust stability with respect to constant real parameters is equivalent to robust stability with respect to time-varying parameters. So, we consider the case in which

$$\rho_Q < \overline{p} < \rho_G. \tag{7}$$

First we have the following observations.

i) Exponential stability is not guaranteed for arbitrary time-varying $p(\cdot)$ because $\overline{p} > \rho_Q$, and the system *is not* quadratically stable.

ii) Exponential stability is guaranteed for constant p because $\overline{p} < \rho_G$.

iii) For a given $\overline{p}$, exponential stability is guaranteed for time-varying $p(\cdot)$ which are slowly varying in time ([15] and [4]).

Let $\mathcal{C}_{\overline{p},\overline{h}}$ be the class of continuous real vector-valued functions defined on $[0, \infty)$ satisfying (6) and for $i = 1, \ldots, q$ and for $t \in [0, +\infty)$

$$\max\left\{\left|\lim_{\Delta t \to 0^-} \frac{p_i(t + \Delta t) - p_i(t)}{\Delta t}\right|, \left|\lim_{\Delta t \to 0^+} \frac{p_i(t + \Delta t) - p_i(t)}{\Delta t}\right|\right\} \leq \overline{h}. \tag{8}$$

Note that, although $p(\cdot)$ is not guaranteed to be differentiable everywhere, we assume that, for all t, the left and right limit of $\dot{p}(t)$ belong to the hypercube $\mathcal{S}_{\overline{h}}$, where

$$\mathcal{S}_{\overline{h}} := [-\overline{h}, \overline{h}]^q \subset \mathbb{R}^q$$

We denote the vertices of $\mathcal{S}_{\overline{h}}$ by $h_{(j)} = (\, h_{1(j)} \quad h_{2(j)} \quad \ldots \quad h_{q(j)}\,)^T$, $j = 1, \ldots, 2^q$.

We consider the following two problems.

Problem 1 Given system (1) with $p(\cdot)$ being a vector of *unknown* time-varying parameters belonging to the class of functions $\mathcal{C}_{\overline{p},\overline{h}}$, where $\rho_Q < \overline{p} < \rho_G$, *establish* a condition for the uniform exponential stability of system (1) for all possible $p(\cdot) \in \mathcal{C}_{\overline{p},\overline{h}}$.

Problem 2 Given $\rho_Q < \overline{p} < \rho_G$, define

$$\overline{h}_{max} := \inf\left\{\overline{h} > 0 \;:\; \text{exists}\, p(\cdot) \in \mathcal{C}_{\overline{p},\overline{h}} \,\text{such that system (1) is not exponentially stable}\right\};$$

then *find* a lower bound $\overline{h}_{lb}$ on $\overline{h}_{max}$. ∎

In this paper we are interested in Problem 2, whose solution gives a lower bound $\overline{h}_{lb}$ on the stability margin $\overline{h}_{max}$ defined to be the supremal allowable rate of variation of parameters which guarantees a stable closed loop system [1]. However our effort will be devoted to solving the first problem; the latter problem can be reduced to a simple binary search algorithm based on the solution of Problem 1.

Following the guidelines of [1], [7], [6], [18], [5] we propose a methodology based on the use of parameter dependent Lyapunov functions; our starting point is the next lemma.

Condition 1 For all $p \in \mathcal{S}_{\overline{p}}$

$$A^T(p)P(p) + P(p)A(p) + \sum_{i=1}^{q} \frac{\partial P}{\partial p_i}(p) h_{i(j)} < 0 \qquad \text{for } j = 1, 2, \ldots, 2^q \tag{9}$$

■

Lemma 1 *Assume there exists a continuously differentiable, simmetric matrix valued function* $P(\cdot) : \mathcal{S}_{\overline{p}} \mapsto \mathrm{IR}^{n\times n}$ *which satisfies Condition 1. Then system (1) is uniformly exponentially stable for all time-varying parameters* $p(\cdot) \in \mathcal{C}_{\overline{p},\overline{h}}$.

Proof. This lemma has been proved in [2] using the Lyapunov function

$$v(x, p) = x^T P(p) x \tag{10}$$

and evaluating the Dini derivative of v along the solutions of system (1). Following arguments similar to [18] it can be shown that condition (9) also guarantees positive definiteness of $P(\cdot)$. ■

Lemma 1 is an answer to Problem 1; now the question which arises quite naturally concerns the choice of a structure for $P(\cdot)$. An answer will be given in the next section.

3 Solution of Problem 1

In previous papers the structure for $P(\cdot)$ was chosen to be affine (that is $P(\cdot)$ mimicked the dependence of the system matrix on p); the coefficients of $P(\cdot)$ were left as degree of freedom for optimization in order to find, if existing, an optimal parameter dependent Lyapunov function which satisfies the hypothesis of Lemma 1.

This approach, although leading to elegant sufficient conditions for the exponential stability of system (1) in terms of LMIs, may render the same conditions quite conservative, as we will show later in the paper by an example.

At the moment, to give a rough idea of such conservatism, note that, since condition (9) implies that for all $p \in \mathcal{S}_{\overline{p}}$ and for all $h \in \mathcal{S}_{\overline{h}}$

$$A^T(p)P(p) + P(p)A(p) + \sum_{i=1}^{q} \frac{\partial P}{\partial p_i}(p) h_i < 0\,, \tag{11}$$

we have, rewriting (11) for $h = 0$, that for all $p \in \mathcal{S}_{\overline{p}}$

$$A^T(p)P(p) + P(p)A(p) < 0 \tag{12}$$

---

[1] Note that $\overline{h}_{max}$ is a strictly positive number.

Hence condition (12) is necessary for (9). Now *it is not* guaranteed, in general, the existence of an affine $P(\cdot)$ satisfying (12) (except that for the one parameter case). Therefore, in some cases, the approach of previous papers may return a *zero* value for $\overline{h}_{lb}$ when solving Problem 2.

The key point is that in previous papers the information concerning the fact that *system (1) is asymptotically stable for constant parameters* (recall that $\overline{p} < \rho_G$) has not been used. This information allows us to construct a family of positive definite $P(\cdot)$, such that the left hand side in (12) be negative definite, just observing that for all positive definite matrix valued function $Q(p)$ and for all $p \in \mathcal{S}_{\overline{p}}$ the Lyapunov equation

$$A^T(p)P(p) + P(p)A(p) = -Q(p) \tag{13}$$

admits a unique positive definite solution $P(p)$ which obviously satisfies (12). In short, the novel idea is that of *optimizing over the set of the* $Q(\cdot)$*'s to obtain suitable* $P(\cdot)$*'s.* We will further discuss this point later, after the proof of Theorem 1.

First we need some preliminary notation. The symbol $M \otimes N$ denotes the Kronecker product between the matrices M and N while the expression

$$X = \mathrm{Lyap}(M, N, L)$$

where $M \in \mathbb{R}^{m\times m}$ and $N \in \mathbb{R}^{n\times n}$, means that $X \in \mathbb{R}^{m\times n}$ is the solution of the (generalized) Lyapunov equation

$$MX + XN = -L \tag{14}$$

Recall that equation (14) admits a unique solution X if and only if $\lambda_i(M)+\lambda_j(N) \neq 0$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, (see [9], p. 348). By I_n we denote the identity matrix of order n. Finally, according to [13], given $F(\cdot) : \mathbb{R}^q \to \mathbb{R}^{n\times m}$, $p \mapsto F(p)$, we define the *Vetter* derivative of F with respect to p as follows

$$\frac{dF}{dp} := \begin{pmatrix} \frac{dF}{dp_1} & \frac{dF}{dp_2} & \cdots & \frac{dF}{dp_q} \end{pmatrix}^T$$

It is fundamental the following theorem.

Theorem 1 *Assume there exists a symmetric, continuously differentiable matrix valued function* $Q(\cdot) : \mathcal{S}_{\overline{p}} \to \mathbb{R}^{n\times n}$ *such that for all* $p \in \mathcal{S}_{\overline{p}}$ *and for* $j = 1, \ldots, 2^q$

$$-Q(p) + X(p)(h_{(j)} \otimes I_n) < 0 , \tag{15}$$

where

$$X(p) = \mathrm{Lyap}\left(A^T(p), I_q \otimes A(p), \left(\frac{dQ}{dp}(p)\right)^T + \left(\frac{dA}{dp}(p)\right)^T \left(I_q \otimes \mathrm{Lyap}(A^T(p), A(p), Q(p))\right)\right.$$
$$\left. + \mathrm{Lyap}(A^T(p), A(p), Q(p)) \left(\frac{dA^T}{dp}(p)\right)^T\right) , \tag{16}$$

then there exists a symmetric, continuously differentiable matrix valued function $P(\cdot)$, *defined for all* $p \in \mathcal{S}_{\overline{p}}$ *as the solution of the Lyapunov equation*

$$A^T(p)P(p) + P(p)A(p) = -Q(p) , \tag{17}$$

which satisfies Condition 1 (which in turn, by virtue of Lemma 1, guarantees that system (1) is uniformly exponentially stable for all time-varying parameters $p(\cdot) \in \mathcal{C}_{\overline{p},\overline{h}}$*).*

In order to optimize over the set of the $Q(\cdot)$'s we need to fix a structure for $Q(p)$: we concentrate on matrix functions $Q(\cdot)$ which are infinitely differentiable. The Mac Laurin series expansion of a matrix function $\tilde{Q}(\cdot)$, if existing, which satisfies the hypothesis of Theorem 1 has the following expression

$$\tilde{Q}(p) = \tilde{Q}(0) + \sum_{i=1}^{\infty} \frac{1}{i!} \left(\frac{\partial \tilde{Q}}{\partial p_1}(0) p_1 + \cdots + \frac{\partial \tilde{Q}}{\partial p_q}(0) p_q \right)^{(i)} \tag{18}$$

where the symbol $(\cdot)^{(i)}$ denotes the "symbolic power" of the argument; for example if $q = 2$ and $i = 2$ we have

$$\left(\frac{\partial \tilde{Q}}{\partial p_1}(0) p_1 + \frac{\partial \tilde{Q}}{\partial p_2}(0) p_2 \right)^{(2)} = \frac{\partial^2 \tilde{Q}}{\partial p_1^2}(0) p_1^2 + 2 \frac{\partial^2 \tilde{Q}}{\partial p_1 \partial p_2}(0) p_1 p_2 + \frac{\partial^2 \tilde{Q}}{\partial p_2^2}(0) p_2^2 \tag{19}$$

Our objective is that of approximating $\tilde{Q}(\cdot)$ as best as possible. Therefore we look for a $Q(\cdot)$ belonging to the following set:

$$\mathcal{Q} = \left\{ Q(\cdot) = Q(\cdot)^T \; : \; Q(p) = \sum_{\substack{i_1 + \cdots + i_q \\ = 0}}^{r} Q_{i_1, \ldots, i_q} p_1^{i_1} p_2^{i_2} \cdots p_q^{i_q} \, , \; Q_{i_1, \ldots, i_q} \in \mathbb{R}^{n \times n} \right\} \tag{20}$$

where the symmetric matrices $Q_{i_1, \ldots, i_q}$ are left free for optimization purposes; such optimization will find the "best" r-order approximation of $\tilde{Q}(\cdot)$. Clearly the value of r (which we would like to choose as large as possible) is limited by the capacity of the computer used to perform the optimization (we will come back on computational burden of our approach later).

Now the problem is that of finding a $Q(\cdot) \in \mathcal{Q}$ which satisfies (15); such problem can be restated as feasibility problem as follows.

Problem 3 Find symmetric matrices $Q_{i_1, \ldots, i_q} \in \mathbb{R}^{n \times n}$, $i_1 + \cdots + i_q \in \{0, 1, \ldots, r\}$, such that

$$Q(p) = \sum_{\substack{i_1 + \cdots + i_q \\ = 0}}^{r} Q_{i_1, \ldots, i_q} p_1^{i_1} p_2^{i_2} \cdots p_q^{i_q} \tag{21}$$

satisfies (15). ∎

Problem 3, as stated, cannot be solved because it involves an infinity of constraints. Therefore we must transform constraints (15) into a finite number of conditions; moreover we have to show that the obtained conditions define a *convex* feasibility problem.

In order to solve Problem 3, we propose the following algorithm which follows the guidelines of the one proposed in [11].

Algorithm 1

Step 1 Define the set

$$\mathcal{P} = \left\{ p^{(i)}, i = 1, \ldots, N \, , \, p^{(i)} \in S_{\overline{p}} \right\}$$

Step 2 Solve the following feasibility problem, subject to a *finite* number of constraints:

Problem 4 Find symmetric matrices $Q_{i_1,\ldots,i_q} \in \mathbb{R}^{n\times n}$, $i_1 + \cdots + i_q \in \{0, 1, \ldots, r\}$, such that

$$- Q(p^{(i)}) + X(p^{(i)})(h_{(j)} \otimes I_n) < 0 \qquad i = 1, \ldots, N\,, \quad j = 1, \ldots, 2^q\,; \tag{22}$$

■

Step 3 If Problem 4 is unfeasible then **STOP**; Problem 1 is unsolvable with the approach of this paper.

If Problem 4 is feasible, given a feasible solution $Q^*(p)$, test that (15) is satisfied for $Q(p) = Q^*(p)$; this is equivalent to check that

$$\max_{\substack{p \in \mathcal{S}_{\overline{p}} \\ j=1,\ldots,2^q}} \lambda_{max} \left(-Q^*(p) + X^*(p)(h_{(j)} \otimes I_n) \right) < 0 \tag{23}$$

where $X^*(p)$ denotes the matrix valued function $X(p)$ evaluated in correspondence of $Q^*(p)$. If (23) is satisfied then **STOP**; we have solved Problem 1. Otherwise goto Step 4.

Step 4 Augment the set $\mathcal{P}$ by introducing (some of) the points where (23) is not satisfied. Define the augmented set as $\mathcal{P}'$; then set $\mathcal{P} = \mathcal{P}'$ and goto Step 1.

Theorem 2 *The constraints in (22) define a set of LMIs in the variables* $Q_{i_1,\ldots,i_q}$.

A number of comments is in order.

Remark 1 The importance of Theorem 1 relies in the fact that it allows to evaluate, for a given p, the derivative of $P(p)$ (namely $X(p)$ in (15)) *without having explicit knowledge of the analytical expression of such derivative*, which results to be very cumbersome also for low order systems. Indeed, to obtain such expression, one should first compute $P(p)$ using the formula (see [13])

$$\mathrm{vec}[P(p)] = -\left(A^T(p) \otimes I_n + I_n \otimes A^T(p) \right)^{-1} \mathrm{vec}[Q(p)]\,, \tag{24}$$

where vec$[F]$ denotes the operation of stacking the columns of the matrix F, (note that the matrix to be inverted at the right hand side in (24) is of order $n^2 \times n^2$), and then compute the Vetter derivative of such matrix; both this computations should have to be performed using symbolic calculus. ■

Remark 2 Theorem 2 guarantees that Problem 4 can be solved using commercially available software [19]. ■

Remark 3 The existence of a solution of Problem 4 is only necessary for (15). Therefore we need to check sufficiency in Step 3; since the form at the left hand side in (15) does not exhibit convexity properties, condition (23) must be checked via force gridding. ■

Remark 4 In principle Algorithm 1 terminates when either Problem 4 is unfeasible or when the check in Step 3 is satisfied. Actually, since the increment of the set $\mathcal{P}$ renders Problem 4 harder and harder to solve, we could reach the situation in which, due to numerical difficulties, it becomes impossible to perform Step 2 of the algorithm. In that case we conclude, empirically, that Problem 1 does not admits a solution with the approach pursued in this paper. ■

Remark 5 Here we discuss computational complexity of Algorithm 1. First in Step 2 we have to solve a LMIs based optimization problem with $N(r,q)$ symmetric matrix variables (where, for instance, $N(r,1) = r+1$, $N(r,2) = \sum_{i=0}^{r}(i+1)$, etc.). Obviously we would like to choose r as big as possible; on the other hand the choice of r is limited by both the order of the system n and the number of parameters q. Concerning Step 3 of Algorithm 1, we need to evaluate the maximum eigenvalue of a symmetric matrix performing a force gridding on a hyper-rectangle. We need to take a sufficient number of points, such to have practical certainty that the test is satisfied. Clearly this requires a lot of computer calculations which, taking into account the fact that we are facing with an off-line problem and the speed of modern computers, renders this technique applicable to systems containing up to 3 parameters. ■

Example 1 Consider system (1) with

$$A_0 = \begin{pmatrix} 0 & 1 \\ -10 & -10 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \qquad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \tag{25}$$

and $\overline{p} = 9.5$.

It is readily seen that this system is asymptotically stable for constant parameters; moreover the system is not quadratically stable.

We have solved Problem 2 choosing for $Q(\cdot)$ the structure (20) with $r = 6$; we found $\overline{h}_{lb} = 1.85$. Then we solved Problem 2 by using affine parameter dependent Lyapunov functions; we found $\overline{h}_{lb} = 0.4$.

This clearly shows that our method obtains a strong improvement with respect to previous ones. Also this example shows that the common idea of choosing for $P(\cdot)$ a structure which mimics that one of the system may lead to overly conservative results.

References

[1] Amato F., *Stability Analysis of Linear Systems Depending on Uncertain and Time-Varying Parameters*, PHD Thesis, Università di Napoli, Feb. 1994.

[2] Amato F., Celentano G., "Robust stability analysis for uncertain linear systems subject to time-varying parameters," *Proceedings of the 4th Mediterranean Symposium in Control and Automation*, Krete, June 1996.

[3] Amato F., Celentano G., "A new approach to the synthesis of parameter dependent Lyapunov functions for the estimation of the stability margin in the presence of slowly varying parameters", *Internal Report*, University of Naples, July 1997.

[4] Amato F., G. Celentano, F. Garofalo, "New sufficient conditions for the stability of slowly-varying linear systems," *IEEE Trans. Aut. Contr.*, Vol. AC-38, pp.1409-1411, Sept. 1993.

[5] Amato F., M. Corless, M. Mattei, R. Setola, "A robust stability margin in the presence of time varying, bounded rate gains," *Int. J. Robust and Nonlinear Control*, Vol 7, pp. 127–143, Feb. 97.

[6] Amato F., M. Corless, M. Mattei, R. Setola, "Robust stability of linear systems depending on Uncertain, slowly-varying parameters," *In Proc. 33rd Allerton Conference*, Monticello (IL), Oct. 1995.

[7] Amato F., A. Pironti, R. Setola, "Robust stability of closed loop SISO linear systems depending on uncertain and time-varying gains." *In Proc. 13th Modelling, Identification and Control*, Grindelwald (Switzerland), Feb. 1994.

[8] Amato F., F. Garofalo, L. Glielmo, A. Pironti, "Robust and quadratic stability via polytopic set covering," *Int. J. Robust and Nonlinear Control*, Vol. 5, pp. 745–756, 1995.

[9] Anderson B. D. O., J. B. Moore, *Optimal Control*, Prentice Hall, Englewood Cliffs, 1989.

[10] Barmish B. R., " Stabilization of uncertain systems via linear control," *IEEE Trans. Aut. Contr.*, vol. AC-28, pp. 848–850, 1983.

[11] Becker G., P. Bendotti, P. Gahinet, C. M. Falinower, "Analysis and controller synthesys for a pressurized water reactor using linear parameter varying systems, *Proceedings of the 35th Conference on Decision and Control*, Kobe, Dec. 1996.

[12] Boyd S., L. El Ghaoui, E. Feron, V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Press, 1994.

[13] Brewer J. W., "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circ. Syst.*, Vol. CAS-25, No. 9, pp. 772–781, Sept. 1978.

[14] Corless M., "Robust stability analysis and controller design with quadratic Lyapunov Functions," *In Variable Structure and Lyapunov Control*, A. S. I. Zinober (editor), Springer Verlag, Berlin, 1993.

[15] Desoer C. A., "Slowly varying systems $\dot{x} = A(t)x$." *IEEE Trans. Aut. Contr.*, vol. AC-14, pp. 780–781, 1969.

[16] De Gaston R. R. E., M. G. Safonov, "Exact calculation of the multivariable stability margin," *IEEE Trans. Aut. Contr.*, vol. AC-33, pp. 156–171, 1988.

[17] Feron E., P. Apkarian, P. Gahinet, "Analysis and synthesis of robust control systems via parameter-dependent Lyapunov functions," *IEEE Trans. Auto. Contr.*, AC-41, pp. 1041–1046, July 1996.

[18] Gahinet P., P. Apkarian, M. Chilali, "Affine parameter dependent Lyapunov functions and real parametric uncertainty," *IEEE Trans. Auto. Control*, AC-41, pp. 436–442, Mar. 1996.

[19] Gahinet P., A.Nemirowski, A.J.Laub, M.Chilali, *LMI Control Toolbox*, The Mathworks, Natick (MA), 1995.

Factor Graphs and Algorithms

Brendan J. Frey
The Beckman Institute
405 North Mathews Avenue
Urbana, IL 61801, USA
`frey@cs.utoronto.ca`

Frank R. Kschischang
ECE Department
University of Toronto
Toronto, Ontario M5S 3G4, CANADA
`frank@comm.utoronto.ca`

Hans-Andrea Loeliger
Endora Tech AG
Gartenstraße 120
CH-4052 Basel, SWITZERLAND
`haloeliger@access.ch`

Niclas Wiberg
Dept. of Electrical Engineering
Linköping University
S-581 83 Linköping, SWEDEN
`nicwi@isy.liu.se`

Abstract—A factor graph is a bipartite graph that expresses how a global function of several variables factors into a product of local functions. Factor graphs subsume many other graphical models, including Bayesian networks, Markov random fields, and Tanner graphs. We describe a general algorithm for computing "marginals" of the global function by distributed message-passing in the corresponding factor graph. A wide variety of algorithms developed in the artificial intelligence, statistics, signal processing, and digital communications communities can be derived as specific instances of this general algorithm, including Pearl's "belief propagation" and "belief revision" algorithms, the fast Fourier transform, the Viterbi algorithm, the forward/backward algorithm, and the iterative "turbo" decoding algorithm.

1 Introduction

A *factor graph* is a bipartite graph that expresses how a global function of several variables factors into a product of local functions.

Example 1. Suppose the real-valued function $g(x_1, x_2, \ldots, x_5)$ of five variables can be written as

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4, x_5) f_2(x_2) f_4(x_4).$$

We refer to g as the *global function* and f_A, f_B, f_C, f_2 and f_4 as *local functions.* The set of arguments of each local function is a subset of the arguments of g.

This factorization can be expressed via the factor graphs shown in Fig. 1. A factor graph consists of two types of vertices: those associated with variables (the unfilled circles in Fig. 1, called *variable nodes*) and those associated with local functions (the filled circles in Fig. 1, called *subset nodes*). The edges of the factor graph are precisely those that join the variable node for x_i to the subset node for f if and only if x_i is an argument of f.

In general, let $X = \{x_i\}_{i \in N}$ be a collection of variables, indexed by a finite set $N = \{1, 2, 3, \ldots, n\}$. If E is a nonempty subset of N, we denote by X_E the subset of X

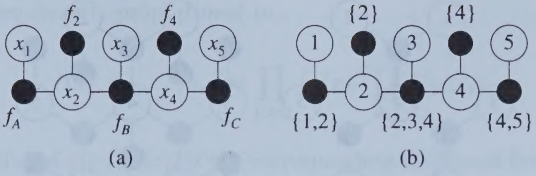


Figure 1: A factor graph that expresses that a global function factors as the product of local functions $f_A(x_1, x_2)f_B(x_2, x_3, x_4)f_C(x_4, x_5)f_2(x_2)f_4(x_4)$. (a) Variable/function form, (b) index/subset form.

indexed by E. For each $i \in N$ the variable x_i takes on values from the *alphabet* A_i. In this paper, we will take A_i to be finite for all $i \in N$, though this assumption is not essential. A particular assignment of a value to each of the variables of X will be referred to as a *configuration* of the variables. Configurations of the variables can be viewed as being elements of the Cartesian product $W = \prod_{i \in N} A_i$, called the *configuration space*. An element $w = (w_1, w_2, \dots, w_n) \in W$, with $w_i \in A_i$ is equivalent to the variable assignment $x_1 = w_1$, $x_2 = w_2$, ..., $x_n = w_n$, and vice versa. We will have occasion to view configurations both as assignments of values to variables, and as elements of W.

We are interested in functions with the elements of X as arguments, i.e., functions with domain W. Let $g : W \to R$ denote such a function, referred to here as the *global function*. For the moment we take the codomain, R, of g to be the set of real numbers, though later, following [1–4], we shall allow R to be any semiring.

Let Q be a collection of subsets of N (i.e., a subset of the power set of N), not including the empty set. Suppose that g can be written as a product of local functions with arguments indexed by the elements of Q, i.e.,

$$g(X) = \prod_{E \in Q} f_E(X_E). \tag{1}$$

Then a factor graph representation of (1) is a bipartite graph with vertex set $N \cup Q$, and edge set $\{\{i, E\} : i \in N, E \in Q, i \in E\}$. As stated earlier, we refer to those vertices that are elements of N as *variable nodes* and those vertices that are elements of Q as *subset nodes*. An edge joins a variable node i to a subset node E if and only if $i \in E$, hence the factor graph is a graphical representation of the relation "element of" in $N \times E$. In Example 1, we have $N = \{1, 2, 3, 4, 5\}$, and $Q = \{\{1, 2\}, \{2, 3, 4\}, \{4, 5\}, \{2\}, \{4\}\}$.

2 Examples of Factor Graphs

2.1 Set Membership Indicator Functions

In coding theory, as in systems theory, one is often interested in describing *subsets* B of the set W of possible configurations. In the coding context, such a subset defines a *code*, while in systems theory, B is referred to as the system's *behavior*, and each element of B is a *valid* configuration.

Set membership can in general be described with a binary-valued global indicator func-

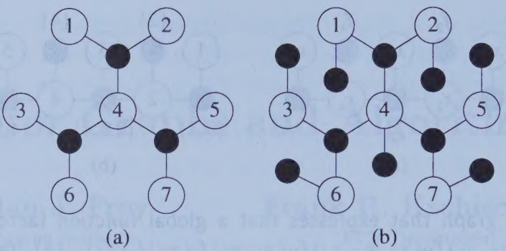


Figure 2: Factor graph representation of (a) a binary code, and (b) the *a posteriori* joint probability mass function of the codeword symbols.

tion. For a set $B \subset W$, define $I_B : W \to \{0, 1\}$ by

$$I_B(w) = \begin{cases} 1 & \text{if } w \in B; \\ 0 & \text{if } w \notin B. \end{cases}$$

It is often the case—by circumstance or by design—that this global indicator function factors into a product of *local* indicator functions, each of which checks for locally valid behavior. Specifically, it may happen that $g(X)$ factors as in (1), where each factor $f_E(X_E)$ of $g(X)$ is itself a binary-valued indicator function. A configuration $w \in W$ is said to be locally valid at E if the local function $f_E(w_E)$ evaluates to unity at the restriction w_E of w to E. Thus a configuration $w \in W$ is a valid configuration if and only if it is locally valid at all $E \in Q$.

For example, every binary linear block code can be described by a set of parity-check equations, in which each equation imposes the condition on a codeword $x = (x_1, x_2, \ldots, x_n)$ that its restriction x_E to a subset E of symbol positions must have even parity; i.e., summing in GF(2), $\sum_{i \in E} x_i = 0$.

Example 2. The factor graph corresponding to the binary code with parity-check matrix

$$H = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

is shown in Fig. 2(a).

More generally, one could impose more complicated constraints, for example requiring that x_E be an element of some more complicated linear code. Such codes were considered by Tanner [5]. Even more generally, following Wiberg, *et al.* [3, 6], one could introduce (unobserved) state variables, not considered part of the "codeword," but which participate in defining locally valid behavior. These "local check" descriptions are naturally described using a bipartite factor graph (called a Tanner graph in [3, 6] or, to emphasize that unobserved state variables are permitted, a Tanner-Wiberg-Loeliger graph [7]).

2.2 Indicator Functions and a Posteriori Probabilities

Continuing with Example 2, let us select, with uniform probability, a codeword $(x_1, \ldots, x_n)$ to transmit over a memoryless channel, and suppose that $y = (y_1, \ldots, y_n)$ is the corresponding observed channel output. The *a posteriori* joint probability distribution of

$\{x_1, \ldots, x_n\}$ is then linearly proportional to

$$f(x_1, x_2, \ldots, x_n) = \prod_{E \in Q} f_E(x_E) \prod_{i=1}^{n} f(y_i|x_i), \tag{2}$$

where, for each value of x_i, $f(y_i|x_i)$ is the corresponding likelihood function evaluated at the observed channel output. We observe that (2) admits a factor graph representation obtained by augmenting the factor graph representation of the code itself with "singleton" local functions $f(y_i|x_i)$ connected by a single edge to each variable. This is shown for our example code in Fig. 2(b).

As we shall see, in computations involving the factor graph, such "singleton" local function nodes can be absorbed into the corresponding variable node, so that effectively, the factor graph describing the *a posteriori* joint probability distribution, given the observed channel output, is equivalent to the factor graph of the code itself.

2.3 Markov Random Fields

A Markov random field (see, e.g., [8]) is a graphical model based on an undirected graph $G = (V, E)$ in which each vertex corresponds to a random variable. Denote by $n(v)$ the neighbors of $v \in V$, i.e., the set of vertices of V connected to v by a single edge of E. The graph G is a *Markov random field* (MRF) if the distribution $p(v_1, \ldots, v_n)$ satisfies the local Markov property: $(\forall v \in V)\ \ p(v|V \backslash \{v\}) = p(v|n(v))$. In other words, G is an MRF if every variable v is independent of non-neighboring variables in the graph, given the values of its immediate neighbors. MRFs are well developed in statistics, and have been used in a variety of applications (see, e.g., [8–11]). See [12] for a brief discussion of the use of MRFs to describe codes.

Under fairly general conditions (e.g., positivity of the joint probability density is sufficient), the joint probability mass function of an MRF can be expressed in terms of a collection of Gibbs potential functions, defined on the set Q of maximal cliques in the MRF. (A clique is a collection of vertices which are all pairwise neighbors, and such a clique is maximal if it is not properly contained in any other clique.) In other words, the distribution factors as

$$p(v_1, v_2, \ldots, v_N) = Z^{-1} \prod_{E \in Q} \psi_E(V_E) \tag{3}$$

where Z^{-1} is a normalizing constant. For example (cf. Example 1), the MRF in Fig. 3(a) can be used to express the factorization

$$p(v_1, v_2, v_3, v_4, v_5) = Z^{-1} \psi_A(v_1, v_2) \psi_B(v_2, v_3, v_4) \psi(v_4, v_5).$$

Clearly (3) has precisely the structure needed for a factor graph representation. Indeed, a factor graph representation may be preferable to an MRF in expressing such a factorization, since distinct factorizations, i.e., factorizations with different Qs in (3), may yield precisely the *same* underlying MRF graph, whereas they will always yield distinct factor graphs. (An example in a coding context of this MRF ambiguity is given in [12].)

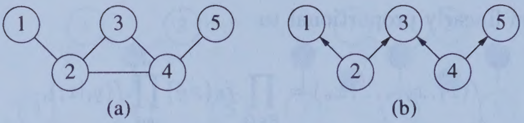


Figure 3: Other graphical models: (a) a Markov random field, and (b) a Bayesian network.

2.4 Bayesian Networks

Bayesian networks (see, e.g., [13–16]) are graphical models for a collection of random variables that are based on directed acyclic graphs (DAGs). Bayesian networks, combined with Pearl's "belief propagation algorithm" [13] have become an important tool in expert systems over the past decade. The first to connect Bayesian networks and belief propagation with applications in coding theory were MacKay and Neal [17], who independently re-discovered Gallager's earlier work on low-density parity-check codes [18] (including Gallager's decoding algorithm.) More recently, at least two papers [12, 19] develop a view of the "turbo decoding" algorithm [20] as an instance of probability propagation in a Bayesian network code model.

Each vertex v in a Bayesian network is associated with a random variable. Denoting by $\mathbf{a}(v)$ the set of *parents* of v (i.e., the set of vertices *from* which an edge is incident on v), the distribution represented by the Bayesian network assumes the form

$$p(v_1, v_2, \ldots, v_n) = \prod_{i=1}^{n} p(v_i|\mathbf{a}(v_i)), \tag{4}$$

where, if $\mathbf{a}(v_i) = \varnothing$, (i.e., v_i has no parents) then we take $p(v_i|\varnothing) = p(v_i)$. For example (cf. Example 1), Fig. 3(b) shows a Bayesian network that expresses the factorization

$$p(v_1, v_2, v_3, v_4, v_5) = p(v_1|v_2)p(v_2)p(v_3|v_2, v_4)p(v_4)p(v_5|v_4).$$

Again, as in the Markov random field case, this graphical model expresses a factorization that is suitable for a factor graph representation.

3 The Sum-Product Algorithm

As in the previous section, let $g(X)$ be a global function over the variables in the set $X = \{x_i : i \in N\}$, with variable x_i taking on values in the finite set A_i. For the moment, we again take g as being real-valued.

In this section we describe a general algorithm that can be used to compute the *marginal functions*

$$G_i(x_i) \triangleq \sum_{x_1 \in A_1, \ldots, x_{i-1} \in A_{i-1}, x_{i+1} \in A_{i+1}, \ldots, x_n \in A_n} g(x_1, \ldots, x_n) \tag{5}$$

for all variables x_i, $i \in N$. We adopt the convention that $\sum_{x_i} f(x_i) = \sum_{x_i \in A_i} f(x_i)$. Similarly, for a subset $J \subset N$, the notation $\sum_{x_i : i \in J} f(X_J)$ means that we sum over all possible configurations of the variables indexed by J. Thus $G_i(x_i) = \sum_{x_j : j \in N \setminus \{i\}} g(X)$.

The definition of marginal function can be extended an arbitrary subset J of N by defining

$$G_J(X_J) \triangleq \sum_{x_i : i \in N \setminus J} g(X).$$

If $g(x_1, \ldots, x_n)$ is a probability distribution, then $G_i(x_i)$ is a marginal distribution, and $G_J(X_J)$ is the joint probability distribution of the variables indexed by J.

When n, the number of arguments of g is small, we will sometimes use a modified notation for the marginal functions. We replace an argument x_i of g with a '+' sign to indicate that the corresponding variable is to be summed over, i.e., "marginalized out."

To perform marginalization, we take advantage of the factorization (1) of the global function, as represented by a factor graph. We use two basic properties of the factor graph: (1) products of local functions can be "gathered" along paths in the graph and (2) variables can be represented in "summary" outside of regions of the graph in which the variable is not "involved."

3.1 An Example

Example 3. To explain what we mean by the rather vague terminology of the previous sentence, let us consider the specific case in which the global function has the structure

$$g(x_1, x_2, x_3) = a(x_1)b(x_1, x_2)c(x_2, x_3). \tag{6}$$

This structure arises, for example, when $g(x)$ is the joint probability distribution of three random variables X_1, X_2, X_3 that form the Markov chain $X_1 \to X_2 \to X_3$ given some observation y.

Consider the computation of $g(x_1, +, +) = p(x_1|y)$. We write

$$\begin{aligned} g(x_1, +, +) &= \sum_{x_2} \sum_{x_3} a(x_1)b(x_1, x_2)c(x_2, x_3) \\ &= a(x_1) \underbrace{\sum_{x_2} b(x_1, x_2) \underbrace{\sum_{x_3} c(x_2, x_3)}_{c(x_2,+)}}_{bc(x_1,+,+)}, \end{aligned} \tag{7}$$

where we write $bc(x_1, x_2, x_3)$ for the product $b(x_1, x_2)c(x_2, x_3)$. In (7) we have identified the various factors that need to be computed to obtain $g(x_1, +, +)$. Our primary observation is that $g(x_1, +, +)$ can be computed knowing just $a(x_1)$ and $bc(x_1, +, +)$. The latter factor can be computed knowing just $b(x_1, x_2)$ and $c(x_2, +)$. These products can be "gathered" along the path from variable node x_3 to the variable node x_1 shown in Fig. 4. Note that variable x_3 can be "summarized," i.e., marginalized out, outside of the dashed region labeled 'x_3;' likewise for x_2.

Analyzing the computation of the remaining marginal functions in the same manner, we find that

$$g(+, x_2, +) = c(x_2, +)ab(+, x_2) \tag{8}$$

$$g(+, +, x_3) = \sum_{x_2 \in S_2} ab(+, x_2)c(x_2, x_3) = abc(+, +, x_3). \tag{9}$$

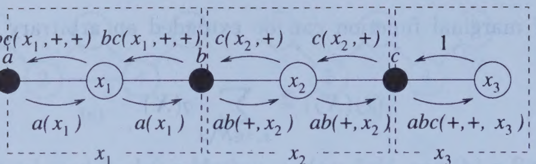


Figure 4: Distributed marginalization for a Markov chain example. For each i, the region in which x_i must be carried is indicated by a dashed box.

Comparing with Fig. 4, we observe that the various factors needed to compute each marginal function can be obtained as products of the "messages" sent along a particular path in the factor graph. The product of the messages carried in the two directions over any given edge is then a marginal function.

3.2 The General Algorithm

The sum-product algorithm operates via a "message passing" procedure that gathers local function products along paths in the factor graph. We assume that this graph is a tree, i.e., that it contains no cycles. The description of the algorithm is simplified by making the assumption that each node of the factor graph is a processor capable of transmitting and receiving "messages" along the edges to which it is connected.

In its simplest realization, the sum-product algorithm operates as follows. The basic operation at each node is to compute the product of the incoming messages at that node. At a subset node, this product is also multiplied by the local function associated with that node. These products are then transmitted on the "outgoing" edges, with the caveat that outgoing messages transmitted along a particular edge should contain no factors received on that edge. In this way, provided that the factor graph contains no cycles, the product of the message *transmitted* along an edge with the message *received* along that edge contains *all* factors of the global function.

This message-passing procedure is initiated at the leaf nodes in the factor-graph, i.e., those nodes on which only a single edge is incident. At leaf nodes that are subset nodes, the transmitted message is a representation of the local function associated with that node. At leaf nodes that are variable nodes, the transmitted message is a "unit" function (the transmission of which, in a practical realization, is of course unnecessary). All other nodes in the graph await the reception of enough messages to be able to compute an outgoing message. Effectively, these nodes wait until messages arrive on all but one edge; when this occurs, the incoming messages are multiplied together with the local function (if any) and the result transmitted on the remaining edge. When a message is *received* on that remaining edge, the appropriate local function products are distributed along all the other incident edges. This general procedure of gathering local function products is illustrated in Fig. 5(a).

The sum-product algorithm is made efficient by observing that the local function products gathered along paths in the factor graph can themselves be marginalized, i.e., not all variables must be *carried* along an edge. In general, a variable needs to be carried if that variable is an argument of a subsequent local function; otherwise, that variable can be marginalized out (cf. Example 3). This marginalization is carried out at the subset

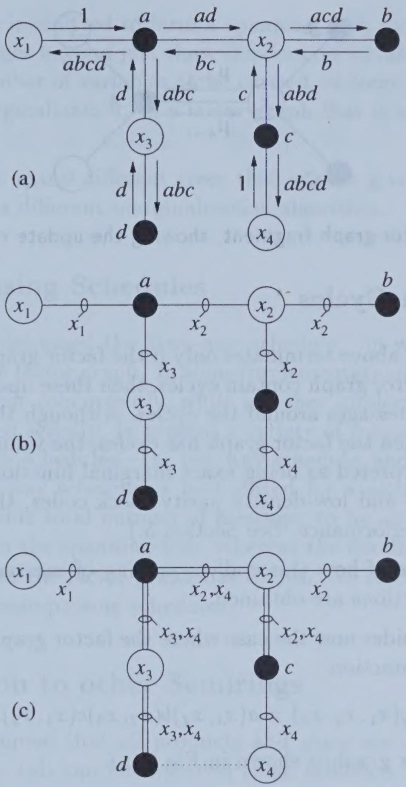


Figure 5: The sum-product algorithm: (a) local function products are "gathered" along paths from the leaves; (b) without cycles, only one variable must be "carried" along each edge; (c) with a cycle, more than one variable must be "carried" on some edges.

nodes, since these nodes are necessarily at the boundaries of the regions defining which variables are to be carried. Fig. 5(b) illustrates which variables are to be carried along each edge for the tree of Fig. 5(a).

Fig. 6 shows a fragment of a specific factor graph, which we assume forms a part of a larger tree. The update rules for this fragment are as follows:

variable to subset: (the product rule)

$$\mu_{x\to A}(x) = \mu_{B\to x}(x) \cdot \mu_{C\to x}(x) \tag{10}$$

subset to variable: (the sum-product rule)

$$\mu_{A\to x}(x) = \sum_{y,z} f_A(x, y, z) \cdot \mu_{y\to A}(y) \cdot \mu_{z\to A}(z). \tag{11}$$

termination:

$$F_x(x) = \mu_{x\to A}(x) \cdot \mu_{A\to x}(x) \tag{12}$$

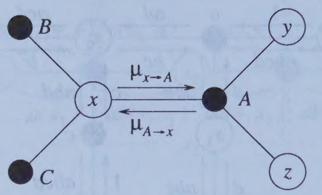


Figure 6: A factor graph fragment, showing the update rules in this case.

3.3 Dealing with Cycles

The algorithm described above terminates only if the factor graph is a tree, i.e., contains no cycles. Should the factor graph contain cycles, then these update rules will lead to the endless propagation of messages around the cycles. Although these messages may "converge" in some sense, when the factor graph has cycles, the values to which the messages converge cannot be interpreted as being exact marginal functions. (Nevertheless, in the decoding of turbo codes and low-density parity-check codes, this "iterative" algorithm *does* provide excellent performance. See Section 5.)

We now give an example of how this endless looping of messages can be dealt with so that exact marginal functions are obtained.

Example 4. Let us consider now the case where the factor graph contains a single cycle. We consider the global function

$$g(x_1, x_2, x_3) = a(x_1, x_2)b(x_2, x_3)c(x_1, x_3). \tag{13}$$

The corresponding factor graph is shown in Fig. 7(a).

To perform exact marginalization, we construct a tree spanning the factor graph. We define a variable x to be "involved" with every subset node E for which $x \in E$. In general, x will be involved with several subset nodes. We say that x must be "carried" over all edges in the spanning tree that are on a path between the nodes with which x is involved. The variable x must not be marginalized out, except when it must no longer be carried. Provided that this rule is followed, exact marginalization is possible.

Thus, if we form the spanning tree shown in Fig. 7(b) (by cutting the cycle in the factor graph at the location indicated), we see that the variable x_3 must be carried over *all* edges in the tree, but that x_1 and x_2 need only be carried over a single edge, respectively. Fig. 5(c) illustrates this situation for the loopy factor graph that includes the dotted

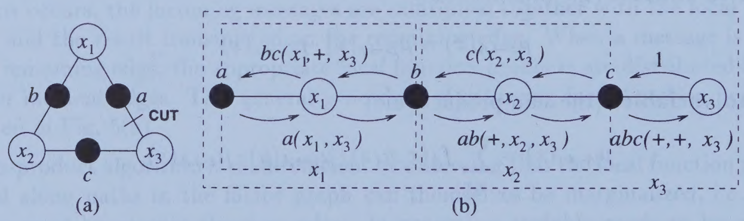


Figure 7: A spanning tree (b) for the factor graph (a), showing that x_3 must be carried everywhere.

edge. If the dotted edge is omitted to form a spanning tree, we see that the variable x_4 must be carried along edges beyond just those incident on variable node x_4. In effect, the "thickness," i.e., the number of variables to be carried, of some edges must be increased to perform an exact marginalization. In a factor graph that is a tree, each edge has unit thickness.

In general, there will be many different trees that span a given "loopy" factor graph, each of which will yield a different marginalization algorithm.

3.4 Message Passing Schedules

Thus far, we have only discussed the "two-way schedule," in which message propagate from the leaf nodes in the factor graph. In some implementations, it may be possible that message passing can occur concurrently, while in other implementations, the messages will of necessity be passed serially. In general a variety of message passing *schedules* are possible. In [12], two such schedules—the two-way schedule, and the flooding schedule—are described. The two-way schedule is best suited for a serial implementation, as it causes the smallest possible total number of messages to be sent; namely, $2E$, where E is the number of edges in the spanning tree, whereas the flooding schedule may lead to a faster convergence when the factor graph has cycles. We refer the reader to [12] for a description of these message-passing schedules.

3.5 Generalization to other Semirings

Up to now, we have assumed that all products and sums are computed in the field of real numbers. Obviously, this can be extended to an arbitrary ring. Indeed, as pointed out by Verdú and Poor [1] and other authors (notably McEliece [2, 4]; see also [3]), the appropriate algebraic structure is that of a semiring, equipped with associative and commutative '+' and '×' operations, and a distributive law that permits distribution of '×' over '+'.

Probably the most relevant such semiring for applications in decoding is the one that replaces '×' with real addition and '+' with 'max' or 'min'. In channel coding for memoryless channels, if we associate with each value that a variable can take on a *cost* equal to the negative log-likelihood of the corresponding channel output, then we can associate with each valid configuration a cost equal to the sum of the costs of the variables. The minimum cost valid configuration is the "maximum-likelihood" sequence, the cost of which is computed (in a factor graph containing no cycles) by the "min-sum" generalization of the "sum-product" algorithm. The "min-sum" formulation is also the framework in which so-called "nonserial" dynamic programming is posed; see [21] for a text book treatment. See [6] for a discussion of the min-sum algorithm.

4 Examples

We now give a few examples of how the sum-product algorithm may be applied in practice.

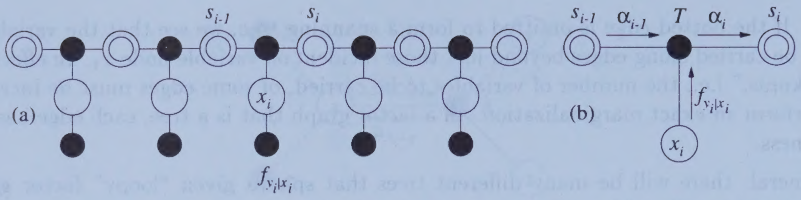


Figure 8: The factor graph for a hidden Markov model (trellis).

4.1 Hidden Markov Models and Trellis Processing

Figure 8(a) shows the factor graph representation for a hidden Markov model. State variables (not observed) are indicated with a double circle. The local function $T(s_{i-1}, x_i, s_i)$ is an indicator function that indicates which state transitions are allowed, and also which outputs (x_i) are associated with the transitions. The output variables x_i are observed at the output of a memoryless channel. Corresponding to symbol x_i, symbol y_i is observed. Since this factor graph is a tree, exact marginalization can be achieved via the sum-product algorithm. This marginalization algorithm is be achieved in two "sweeps" through the graph, running essentially left-to-right (forward) and right-to-left (backward).

Let $\alpha_{i-1}(s_{i-1})$ denote the message sent by state variable s_{i-1} to the local function node T, as shown in Fig. 8(b). Let $f_{y_i|x_i}(x_i)$ denote the message sent by variable node x_i. Then, the sum-product algorithm will compute the message

$$\alpha_i(s_i) = \sum_{x_i} \sum_{s_{i-1}} T(s_{i-1}, x_i, s_i)\alpha_{i-1}(s_{i-1}) f_{y_i|x_i}(x_i)$$

and send this to variable node s_i. This is precisely the "forward" step in the forward/backward algorithm described in [22]. The "backward" (β) step is described in like manner. In this way, the sum-product algorithm specializes, on a trellis, to the well known forward/backward algorithm. It is easy to see that the min-sum version specializes to the Viterbi algorithm (provided we take $T(s_{i-1}, x_i, s_i) = \infty$ for an "invalid" trellis transition). Thus, we recover the well-known trellis-processing algorithms as special cases of the sum-product algorithm.

4.2 The Fast Fourier Transform

Following Aji and McEliece [4], who develop a fast Hadamard transform using a graph-based approach, we now develop the fast Fourier transform (FFT) using factor graphs. For a complex-valued data vector with components $f[n]$, $n \in \{0, 1, \ldots, N-1\}$, the N-point discrete Fourier transform $F[k]$ is defined as

$$F[k] = \sum_{n=0}^{N-1} f[n] W^{nk}$$

where $k \in \{0, 1, \ldots, N-1\}$ and $W = \exp(-j2\pi/N)$ is a complex Nth root of unity.

We consider the case $N = 8$, with $W = \exp(-j\pi/4)$, though the approach we take generalizes in an obvious way to the case where N is an arbitrary power of two. Let

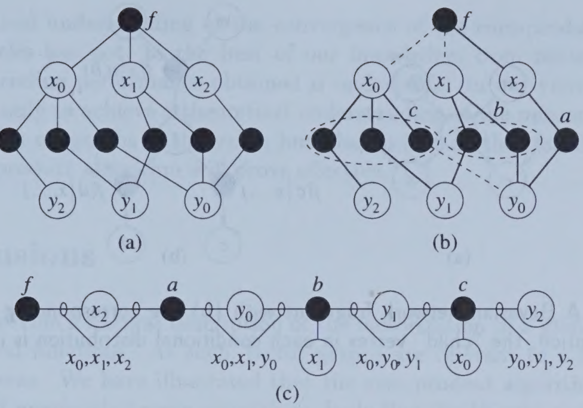

(a) (b) (c)

Figure 9: The discrete Fourier transform kernel: (a) factor graph; (b) a particular spanning tree; (c) spanning tree after clustering, showing which variables must be carried.

$x_0, x_1, x_2, y_0, y_1, y_2$ be binary-valued, and define $n = 4x_2 + 2x_1 + x_0$, $k = 4y_2 + 2y_1 + y_0$. We write the DFT kernel, which we take as our global function, in terms of the x_is and y_is as

$$\begin{aligned} g(x_0, x_1, x_2, y_0, y_1, y_2) &= f[4x_2 + 2x_1 + x_0]W^{(4x_2+2x_1+x_0)(4y_2+2y_1+y_0)} \\ &= f[4x_2 + 2x_1 + x_0](-1)^{x_2y_0}(-1)^{x_1y_1}(-1)^{x_0y_2}(-j)^{x_0y_1}(-j)^{x_1y_0}W^{x_0y_0}, \end{aligned}$$

since $W^{16} = W^8 = 1$, $W^4 = -1$, and $W^2 = -j$. The factor graph corresponding to this global function is shown in Fig. 9(a). We observe that the DFT of $f[n]$ is the marginal function $F[4y_2 + 2y_1 + y_0] = g(+, +, +, y_1, y_2, y_3)$.

The factor graph in Fig. 9(a) has cycles, yet we wish to carry out exact marginalization, so we form a spanning tree. There are many possible spanning trees, of which one is shown in Fig. 9(b). (Different choices for the spanning tree will lead to possibly different DFT algorithms when the min-sum algorithm is applied.) If we cluster the local functions as shown in Fig. 9(b), essentially by defining

$$\begin{aligned} a(x_2, y_0) &= (-1)^{x_2y_0}, \\ b(x_1, y_0, y_1) &= (-1)^{x_1y_1}(-j)^{x_1y_0}, \\ c(x_0, y_0, y_1, y_2) &= (-1)^{x_0y_2}(-j)^{x_0y_1}W^{x_0y_0}, \end{aligned}$$

we arrive at the spanning tree shown in Fig. 9(c). Observe that three binary variables (or eight complex quantities) must be carried over each edge in the path from vertex f to vertex y_2. Along this path, first x_2, then x_1, and then x_0 are marginalized out as y_0, y_1, and y_2 are added to the list of variables to be carried. In three steps, the function $f[n]$ is converted to the function $F[k]$. Clearly we have obtained a fast Fourier transform as an instance of the sum-product algorithm.

4.3 Belief Propagation in Bayesian Networks

As discussed in Section 2.4, the probability distribution (4) corresponding to a Bayesian network has the product form that allows a straightforward conversion to a factor graph

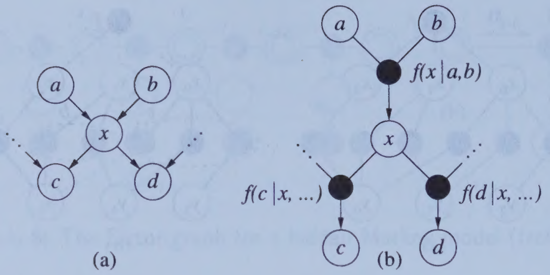


Figure 10: (a) A Bayesian network fragment with (b) the corresponding factor graph, in which, by convention, the "child" vertex in each conditional distribution is indicated with an arrow.

representation. For example, Fig. 10(a) shows a fragment of a Bayesian network, and Fig. 10(b) shows the corresponding factor graph fragment. In general, a local function, corresponding to a single factor of (4), assumes the form $f(x|\mathbf{a}(x))$, where $\mathbf{a}(x)$ is the set of "parents" of variable x in the Bayesian network. In the factor graph representation we have, by convention, identified the "child" vertex by an arrow. This allows the flow of "causality" to be inferred equally well from the factor graph as from the Bayesian network itself. Representing such causality using directed edges is often cited as one of the basic advantages of modeling a probability distribution using a Bayesian network [13]; by including arrows in the corresponding factor graph, this advantage is retained.

It is easy to show that the sum-product algorithm, when applied to cycle-free Bayesian networks, yields Pearl's belief propagation algorithm [13]. See [12] for a discussion of this correspondence.

5 Approximate "Marginalization" in Graphs with Cycles

In a great many applications, one is forced to consider factor graphs that contain cycles. In many such cases, exact marginalization in a spanning tree proves to be computationally infeasible due to the excessive "thickness" of some of the edges. Examples of factor graphs of this type include the graphs that describe low-density parity-check codes introduced by Gallager [18], and the turbo codes introduced by Berrou, *et al.* [20]. (See [3,23] for a Tanner graph description of turbo codes, and [12,19] for a Bayesian network description.)

Nevertheless, by proceeding as if the graph contained no cycles (i.e., by applying the "carry-one-variable" sum-product algorithm), excellent results may be obtained. For example, simulation results of MacKay and Neal [17] indicate that low-density parity-check codes can be decoded (with reasonable complexity) using this approach so that the E_b/N_0 required to achieve a bit error rate of 10^{-5} is within approximately 2dB of the Shannon limit [17] (using antipodal signaling on a white Gaussian noise channel). Even more impressive are the well-known results of Berrou, *et al.*, who show that an iterative sum-product decoding procedure for turbo codes can approach the Shannon limit to within a 0.5 dB in E_b/N_0 required to achieve a bit error rate of 10^{-5}.

While a theoretical understanding of the convergence of the sum-product algorithm in graphs with cycles has not, to the best of our knowledge, been achieved in general, the excellent decoding performance obtained is undeniable. In our view, this is a great motivation, not only to achieve a theoretical understanding of the properties of the sum-product algorithm on graphs with cycles, but also to find further families of codes for which the sum-product algorithm will prove effective.

6 Conclusions

Factor graphs provide a natural description of the factorization of a global function into a product of local functions. As such, factor graphs are relevant to a broad spectrum of application areas. We have illustrated that the sum-product algorithm encompasses a wide variety of previously known algorithms, including the Viterbi algorithm, the forward/backward algorithm, Pearl's belief propagation algorithm, even the fast Fourier transform. No doubt many other algorithms can be captured in this framework. Because of its generality and essential simplicity, we feel that the sum-product algorithm should be included as a part of every engineer's standard "algorithms toolkit."

Acknowledgments

The concept of factor graphs as a generalization of Tanner graphs was devised by a group at ISIT '97 in Ulm that included the authors, G. D. Forney, Jr., R. Kötter, D. J. C. MacKay, R. J. McEliece, and R. M. Tanner. We benefitted greatly from the many discussions on this topic that took place in Ulm.

References

[1] S. Verdú and H. V. Poor, "Abstract dynamic programming models under commutativity conditions," *SIAM J. on Control and Optimization*, vol. 25, pp. 990–1006, July 1987.

[2] R. J. McEliece, "On the BJCR trellis for linear block codes," *IEEE Transactions on Information Theory*, vol. 42, pp. 1072–1092, July 1996.

[3] N. Wiberg, *Codes and Decoding on General Graphs.* PhD thesis, Linköping University, Sweden, 1996.

[4] S. M. Aji and R. J. McEliece, "A general algorithm for distributing information on a graph," in *Proc. 1997 IEEE Int. Symp. on Inform. Theory*, (Ulm, Germany), p. 6, July 1997.

[5] R. M. Tanner, "A recursive approach to low complexity codes," *IEEE Trans. on Inform. Theory*, vol. IT-27, pp. 533–547, Sept. 1981.

[6] N. Wiberg, H.-A. Loeliger, and R. Kötter, "Codes and iterative decoding on general graphs," *European Trans. on Telecommun.*, vol. 6, pp. 513–525, Sep./Oct. 1995.

[7] G. D. Forney, Jr., "The forward-backward algorithm," in *Proc. 34th Annual Allerton Conf. on Communication, Control, and Computing*, (Allerton House, Monticello, Illinois), pp. 432–446, October 1996.

[8] R. Kindermann and J. L. Snell, *Markov Random Fields and their Applications.* Providence, Rhode Island: American Mathematical Society, 1980.

[9] C. J. Preston, *Gibbs States on Countable Sets.* Cambridge University Press, 1974.

[10] V. Isham, "An introduction to spatial point processes and Markov random fields," *Int. Stat. Rev.*, vol. 49, pp. 21–43, 1981.

[11] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in Boltzmann machines," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (D. E. Rumelhart and J. L. McClelland, eds.), vol. I, pp. 282–317, Cambridge MA.: MIT Press, 1986.

[12] F. R. Kschischang and B. J. Frey, "Iterative decoding of compound codes by probability propagation in graphical models," *IEEE J. Selected Areas in Commun.*, vol. 16, Jan. 1998.

[13] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco, CA: Morgan Kaufmann, 1988. Revised second printing.

[14] F. V. Jensen, *An Introduction to Bayesian Networks.* New York: Springer Verlag, 1996.

[15] R. E. Neapolitan, *Probabilistic Reasoning in Expert Systems: Theory and Algorithms.* Toronto: John Wiley & Sons, 1990.

[16] B. J. Frey, *Bayesian Networks for Pattern Classification, Data Compression and Channel Coding.* Toronto, Canada: Department of Electrical and Computer Engineering, University of Toronto, 1997. Doctoral dissertation available at `http://www.cs.utoronto.ca/~frey`.

[17] D. J. C. MacKay and R. M. Neal, "Good codes based on very sparse matrices," in *Cryptography and Coding. 5th IMA Conference* (C. Boyd, ed.), no. 1025 in Lecture Notes in Computer Science, pp. 100–111, Berlin Germany: Springer, 1995.

[18] R. G. Gallager, *Low-Density Parity-Check Codes.* Cambridge, MA: M.I.T. Press, 1963.

[19] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm," *IEEE J. on Selected Areas in Commun.*, vol. 16, Jan. 1998.

[20] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," in *Proc. IEEE Int. Conf. Commun. (ICC)*, (Geneva, Switzerland), pp. 1064–1070, 1993.

[21] U. Bertelè and F. Brioschi, *Nonserial Dynamic Programming.* New York: Academic Press, 1972.

[22] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. on Inform. Theory*, vol. 20, pp. 284–287, Mar. 1974.

[23] N. Wiberg, H.-A. Loeliger, and R. Kötter, "Codes and iterative decoding on general graphs," *European Transactions on Telecommunications*, vol. 6, pp. 513–525, 1995.

Linear-time binary codes correcting localized erasures

A. BARG and S. ZHOU

Bell Laboratories 2C-375
700 Mountain avenue
Murray Hill NJ 07974

Abstract

We study binary codes correcting localized erasures. In this model the encoder knows coordinates of possible erasures in the received sequence. We present two constructions of codes with $t(1+\epsilon)$ check bits, where t is the maximal possible number of erasures. The simplest of the two constructions requires $O(n/\epsilon^3)$ binary operations for encoding and $O(n/\epsilon^2)$ binary operations for decoding, where n is the block length.

1 Introduction

We consider binary block codes. The transmission model that we study is defined as follows. Let E_2^m be the set of all binary strings of length m. Suppose we need to transmit a message $\boldsymbol{m} \in E_2^m$. The message is encoded in a code sequence $\boldsymbol{x} = (x_1, x_2, \dots x_n)$. Let n be the block length of the code and $\mathcal{N} = \{1, 2, \dots, n\}$ the set of code coordinates. Let $U \subset \mathcal{N}$, $|U| = t$, be an arbitrary subset of coordinates that becomes known to the sender before encoding and may vary from message to message. The sequence $\boldsymbol{y} = (y_1, \dots, y_n)$ received from the channel with erasures localized at U is formed as follows: $y_i = x_i, i \notin U$ and $y_i = x_i$ or $y_i =$ erasure for $i \in U$. In other words, the encoder knows that symbols outside U are never erased and symbols in U may or may not be erased.

The maximal possible number of erasures t is known to the decoder and is a parameter of the code. The decoder receives a sequence in which, of course, it knows which symbols are erased. Otherwise, the decoder has no information. The problem is to construct codes of small redundancy and low complexity that correct localized erasures. By constructing a code we mean defining an encoding mapping ϕ that sends pairs $(\boldsymbol{m}, U) \in E_2^m \times \binom{\mathcal{N}}{t}$ to code sequences $\boldsymbol{x} \in E_2^n$ and a decoding mapping ψ such that $\psi(\phi(\boldsymbol{m}, U)) = \boldsymbol{m}$.

Let $r(n,t)$ be the number of check bits in a code C that corrects t localized erasures. Then

$$t \leq r(n,t),$$

since if t erasures occur and the number of checks is less than t, then some of the information bits are lost and the decoder has no means to recover them.

The motivation of the problem is twofold. On the practical side, the problem is related to packet routing in networks employing the Asynchronous Transfer Mode (ATM) and Internet. In networks with high volume of transmitted data, packets or groups of packets are often is lost or delayed due to traffic congestion or buffer overflows. In Internet, for instance, this causes the receiver to issue a retransmission request. However, in on-line applications such as, for instance, real-time video transmission, retransmission is totally impossible. In this context we face the problem of retrieving the lost data at the receiving end at the expense of having a certain data overhead in the transmission. To alleviate this problem, one of the ways is to treat parts of the message differently by assigning them different priorities. Then, for instance, one could guarantee that at least the most important part of the transmission can reach the receiver. Priority transmission in networks of both types has been widely discussed in the literature, see, e.g., [7], [2].

Previous theoretical work to which our paper is closely related is the study of codes correcting localized errors. This problem was introduced in [4]. Under this model, the definition is mostly the same as given above except that bits in the set U are either transmitted correctly or flipped in the channel. Paper [4] shows that if the code length $n \to \infty$, then the number of check bits in an asymptotically optimal binary code correcting $t = \alpha n$ localized erasures equals $nH_2(\alpha)$. This is interesting from two points of view, first that the size of optimal codes is known exactly, and second that the number of checks is much smaller than needed to correct αn usual (Hamming) errors, which in best known codes equals $nH_2(2\alpha)$.

Similarly, the number of checks needed to correct localized erasures in binary codes is much smaller than in codes correcting usual erasures. Namely, the constructions proposed in this paper require $t(1+\epsilon)$ check bits, where ϵ is an arbitrary positive number, and thus are arbitrarily close to the optimum. On the other hand, to correct $t = \alpha n$ usual (i.e., not localized) erasures one needs codes with distance at least $t+1$. This requires at least $nH_2(\alpha)$ checks. Here H_2 is the entropy function.

Note that for larger alphabets there exist MDS codes that correct t usual erasures with t check symbols. The encoding/decoding time is of order $\mathcal{O}(tn)$. However, an easy consequence of the Singleton bound shows that the size q of the alphabet of an MDS code is large, namely $q \geq \max(t+1, n-t+1)$, where t is the number of erasures.

Note also that it is possible to construct probabilistic encoding/decoding methods that ensure recovery of binary sequences with at most t erasures using $t(1+\epsilon)$ check bits, see [8].

2 Low-Complexity Binary Codes: Recursive Constructions

The idea of constructing low-complexity codes is essentially the same for all transmission models. Namely, the code length n is partitioned into segments of smaller length. On each of these segments one uses an optimal code of smaller length. The process is continued recursively until the length is so small that the encoding and decoding of the optimal code can be implemented by exhaustive search. This general construction varies for different models since optimal codes are constructed in different ways.

This idea was used in [9] to construct codes of length n and size $2^{\frac{1}{4}n}$ correcting a small number $t = \alpha n$ of Hamming errors. The encoding/decoding complexity is of order $\mathcal{O}(n)$ and $\alpha \leq 5 \cdot 10^{-7}$.

Paper [1] introduced a recursive construction of binary codes correcting $t = \alpha n$ localized errors with $n(H_2(\alpha) + \epsilon)$ check bits. The complexity of encoding/decoding is $\mathcal{O}(n^4)$, where the hidden constant depends on ϵ.

Recursive codes over a very large alphabet correcting t usual erasures over a very large alphabet were introduced in [3]. The number of check symbols is $t(1 + \epsilon)$. The complexity of encoding/decoding is $\mathcal{O}(n/\epsilon^4)$. Here ϵ must be small since the construction relies on error-correcting codes from [9].

Yet another related transmission model deals with codes correcting memory defects. Asymptotically optimal codes correcting defects were constructed in in [5], [6]. In this model, the set U of size t consists of two parts, $U = U_0 \cup U_1$. The transmitted sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ is received as $\boldsymbol{y} = (y_1 \ldots, y_n)$, where $y_i = x_i$ for $i \notin U$, $y_i = 1$ for $i \in U_1$, and $y_i = 0$ for $i \in U_0$. Again the set U is known to the encoder. The goal is, given a message $\boldsymbol{m} \in E_2^m$, to construct a code sequence $\boldsymbol{x} \in E_2^n$ such that $x_i = 0$ for $i \in U_0$ and $x_i = 1$ for $i \in U_1$. This code sequence then passes unscathed through the channel. It is easily seen that the minimal number of check bits $n - m$ needed to correct t defects is t. Paper [6] introduced codes that use $t + o(t)$ checks. The complexity of encoding is $\mathcal{O}(n \log^2 n/\epsilon^3)$ and of decoding $\mathcal{O}(n \log n/\epsilon)$. In our work we rely on these constructions to construct codes correcting localized erasures. Using these constructions, we are able to achieve linear encoding/decoding complexity for localized erasures. The redundancy of codes is $t(1 + \epsilon)$, which is not possible for usual erasures with deterministic encoding/decoding algorithms.

3 Constructions

We are aiming at constructing codes that correct $t = \alpha n$ localized erasures with linear encoding/decoding complexity. An obvious observation is that this is possible with linear time error-correcting codes from [9]. In a construction of this type the code length is partitioned into segments of length d. The set of possible erasures is known to the encoder. The idea of the construction would be, before encoding, to apply a permutation from a relatively small set known to both parties so that most blocks of length d contain $\leq \alpha d$ erasures. Then on each of these blocks we apply a linear-time error-correcting code. The final problem is to convey to the decoder the number of the permutation. Though theoretically possible, this construction would work only for small α. Furthermore, the complexity depends linearly on n but also on the fraction of errors corrected by linear-time codes from [9], and turns out to be high.

We propose two recursive constructions of type similar to those discussed in the precious section. The general idea is to treat localized erasures as defects of fixed value, say 0. Optimal codes correcting such defects were constructed in [5]. We use these codes as a basis in our recursion.

More precisely, one of the results in [5] suggests considering codes for a simplified model with defects. Suppose that in the above definition of channel with defects one of the sets, say U_1 is empty, i.e., $U = U_0$, and both the encoder and the decoder know that all the defects have value 0 (this is similar to asymmetric errors studied in coding theory earlier). The result proved in [5] is as follows.

Theorem 3.1 [5] *There exist easily constructible binary codes that correct t defects of fixed value. The redundancy of codes is $t+1$. The encoding time equals $\mathcal{O}(t^3+tn)$ and the decoding time $\mathcal{O}(tn)$.*

We return to the definition in the beginning of the paper. In the constructions below we refer to positions of possible erasures in segments of the block of length n. By this we mean the coordinates of the segments that are contained in U.

CONSTRUCTION 1 This construction relies on the above theorem. It has a number of parameters. First, the block length n is divided into segments of a certain length d. Both encoding and decoding are performed separately for each segment. The goal of the encoder is to incorporate in the code sequence the number of possible erasures in each segment. Note that this number can be greater than the number of actually erased bits, visible to the decoder.

The properties of the construction are summarized in the following theorem.

Theorem 3.2 *There exist easily constructible binary codes that correct t localized erasures. The redundancy of codes is $t(1+\epsilon)$, where $\epsilon > 0$ is arbitrary number. The encoding complexity is $\mathcal{O}(n/\epsilon^4)$. The decoding complexity is $\mathcal{O}(n/\epsilon^2)$.*

Proof (outline): Partition the length n into segments of length d, where $d = \mathcal{O}(\epsilon^{-2})$ is a constant. Let x, $0 \leq x \leq 1$ be a parameter. The encoding algorithm works separately on each segment of length d. Let t_i be the number of possible erasures on the current segment. If $t_i \geq xd$, the segment is not used to transmit any information. The encoder fills it with zeros. (We assume that our code does not contain the all-zero sequence.) Otherwise, length d is partitioned into subsegments of length $\ell = \mathcal{O}(\log d)/(1-x)$. One of these subsegments has at most $\frac{t_i}{d}\ell$ erasures. This subsegment also carries no bits of the message transmitted. It is used to convey to the decoder the number of erasures in each subsegment of the current segment. The remaining part of the check bits of this subsegment is used to transmit its location to the decoder. All subsegments are encoded with codes from Theorem 1.

The decoding algorithm, working on a segment of length d, first replaces erased bits by zeros. If the segment turns out to be all-zero, it is discarded. Otherwise, the decoder locates the subsegment that carries the number of possible erasures. Then each subsegment is decoded according to the algorithm from Theorem 1 and the decoder proceeds to the next segment. □

Thus, this construction can be implemented on-line. The next construction does not have this feature, but in exchange has smaller encoding complexity.

CONSTRUCTION 2 This is a recursive construction that performs encoding and decoding based on the entire length n.

Theorem 3.3 *There exist easily constructible binary codes that correct t localized erasures. The redundancy of codes is $t(1+\epsilon)$, where $\epsilon > 0$ is arbitrary number. The encoding complexity is $\mathcal{O}(n/\epsilon^3)$. The decoding complexity is $\mathcal{O}(n/\epsilon^2)$.*

Proof (outline): Suppose there exist codes of length n' that correct t' localized erasures for all $n' < n$, $t' < t$ that have properties required in the theorem. The construction proceeds recursively and relies on this assumption as on induction hypothesis. It has a number of parameters. First the block length n is partitioned into segments of length βn. One of them has at most $\alpha\beta n$ erasures. Denote this segment by I. The remaining length is partitioned into subsegments of length $d = \mathcal{O}(\epsilon^{-3/2})$. The segment of length βn is not used to transmit any bits of the message. Rather, the encoder records on it the number of possible erasures in each of the remaining subsegments.

The encoding algorithm proceeds as follows. First, all positions of possible erasures are replaced by zeros. Next, the encoder finds a segment I of length βn with relatively few erasures. The remaining subsegments of length d are encoded using codes of Theorem 1. The number of possible erasures on these subsegments is encoded on segment I using the code whose properties are guaranteed by the induction hypothesis. We also have to allow a small number of check bits to convey to the decoder the location of segment I.

The decoder first replaces all erased bits with zeros. Then it finds segment I, and decoding it, finds the number of possible erasures in each of the subsegments of length d. Then it applies the decoding algorithm of Theorem 1 to recover the parts of the message transmitted in these subsegments. □

4 Conclusion

This paper introduces binary codes correcting t localized erasures. The redundancy of codes is $t(1+\epsilon)$ and is thus very close to the minimal possible. The complexity is $\mathcal{O}(n/\epsilon^4)$ or $\mathcal{O}(n/\epsilon^3)$ for encoding and $\mathcal{O}(n/\epsilon^2)$ for decoding. The number ϵ can take any positive value, contrary to some other recursive constructions in which it has to be small, which yields high encoding/decoding complexity. We note that it is possible to construct almost optimal low-complexity binary codes that correct localized errors and erasures at the same time. This will be the subject of a future work. An interesting problem would be to construct codes correcting t localized erasures with redundancy $t + o(n)$ as for codes correcting defects [6].

References

[1] R. Ahlswede, L. A. Bassalygo, and M. S. Pinsker, "Asymptotically optimal binary codes of polynomial complexity correcting localized errors," *Problemy Peredachi Informatsii*, **31** (2) (1995), 76–83 (in Russian) and 162–168 (English translation).

[2] A. Albanese, J. Blömer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoding transmission," *IEEE Trans. Inform. Theory*, **IT-42** (6) (1996), 1737–1747.

[3] N. Alon and M. Luby, "A linear time erasure resilient code with nearly optimal recovery," *IEEE Trans. Inform. Theory*, **IT-42** (6) (1996), 1732–1736.

[4] L. A. Bassalygo, S. I. Gelfand, and M. S. Pinsker, "Codes for channels with localized errors," in: *Proc. 4th Swedish–Soviet Workshop on Information Theory*, (1989), pp. 95–99.

[5] I. I. Dumer, "Asymptotically optimal codes correcting memory defects of fixed multiplicity," *Problemy Peredachi Informatsii*, **25** (4) (1989), 3–10 (in Russian) and 259–264 (English translation).

[6] ——, "Asymptotically optimal linear codes correcting defects of linearly increasing multiplicity," *Problemy Peredachi Informatsii*, **26** (2) (1990), 3–17 (in Russian) and 93–104 (English translation).

[7] A. I. Elwalid and D. Mitra, "Statistical multiplexing with loss priorities in rate-based congestion control of high speed networks," *IEEE Trans. Commun.*, **42** (11) (1994), 2989–3002.

[8] M. G. Luby, M. Mitzenmacher, M. Amin Shokrollahi, D. A. Spielman, and V. Stemann, "Practical loss-resilient codes," *Proc. 29th ACM Annual Sympos. on the Theory of Computing* (STOC'97), ACM (1997), pp. 150–159.

[9] D. Spielman, "Linear-time encodable and decodable error-correcting codes," *IEEE Trans. Inform. Theory*, **IT-42** (6) (1996), 1723–1731.

The convoluted history of minimal encoders

Hans-Andrea Loeliger
Endora Tech AG, Gartenstrasse 120
CH-4052 Basel, Switzerland

email: haloeliger@access.ch

Abstract: Kalman's classic formula *"minimal ⇔ controllable and observable"* fundamentally solves the minimality problem of linear systems theory. However, it took 35 years until Kalman's formula was found to apply also to convolutional encoders.

1 Introduction

A convolutional encoder is a linear sequential circuit as shown in Fig. 1. In mathematical terms, an encoder is described by a set of equations

$$\begin{aligned} \mathbf{x}(j+1) &= A\mathbf{x}(j) + B\mathbf{u}(j), \\ \mathbf{y}(j) &= C\mathbf{x}(j) + D\mathbf{u}(j), \end{aligned} \tag{1}$$

where $j \in Z$ is the discrete time index, where $\mathbf{u}(j) = [u_1(j), \ldots, u_k(j)]$ are the time-j input variables, $\mathbf{y}(j) = [y_1(j), \ldots, y_n(j)]$ are the output variables, $\mathbf{x}(j) = [x_1(j), \ldots, x_m(j)]$ are the state variables, and where A, B, C, and D are matrices of the appropriate dimensions; the equation is over the binary field GF(2) (or over any field).

An encoder as in Fig. 1 or, equivalently, as in (1) is *minimal* if no smaller encoder, (i.e., with a smaller number m of delay cells) produces the same set of possible output sequences. (For an in-depth discussion of minimality, see [1].)

How can we test whether a given encoder is minimal? Kalman's formula [2] [3]

$$\textit{minimal} \Leftrightarrow \textit{controllable and observable} \tag{2}$$

has long been known to give the answer for the *different* notion of minimality in traditional linear systems theory: there, a system (1) (over the real numbers or over any field) is minimal if no such system with a smaller state space dimension m gives the same *transfer function*, i.e., the same set of input/output sequence pairs. In that theory, a system (1) is *controllable* if the block matrix $\mathcal{C} \triangleq [B, AB, \ldots, A^{m-1}B]$ has full rank m; it is *observable* if the block matrix $\mathcal{O} \triangleq [C^T, A^TC^T, \ldots, (A^T)^{m-1}C^T]$ has full rank m. Clearly, the rank test of $\mathcal{C}$ and $\mathcal{O}$ is *not* a minimality test for a convolutional encoder.

2 Detours

The mentioned facts were well understood when the system-theoretic study of convolutional encoders began [4]. However, the subsequent work (most notably [5]) shifted away from (1) and focussed on the sequence equation

$$\mathbf{y}(D) = G(D)\mathbf{u}(D) \tag{3}$$

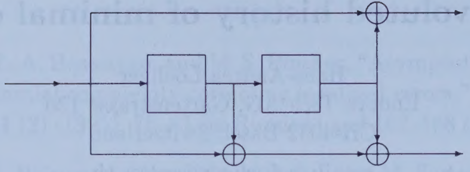


Figure 1: A binary convolutional encoder (with $k = 1$, $n = 2$, and $m = 2$).

where $\mathbf{y}(D)$ and $\mathbf{u}(D)$ are formal Laurent series and where the *generator matrix* (or *encoding matrix*) $G(D)$ is a polynomial or rational matrix in D. Various notions of minimality were defined for such matrices [5] [7], the relation of which to the minimality of (1) is subtle (see [1, Section V.A]). Moreover, the unfortunate custom of referring also to $G(D)$ as "encoder" has confused many students of the subject to the extent that few engineers know how to test whether a general encoder (1) is minimal.

We shall not further use $G(D)$ in this paper.

3 Observability Revisited

A *trellis section* is a four-tuple $X = (G, S, S', B)$, where G is the *label alphabet,* S and S' are the *left state space* and the *right state space,* respectively, and the *branches* B are a subset of $S \times G \times S'$. In this paper, we will assume $S = S'$. A *trellis* is a sequence $\mathcal{X} = \{X_j\}_{j \in Z}$ of trellis sections $X_j = (G_j, S_j, S'_j, B_j)$ such that $S_j = S'_{j-1}$ for all $j \in Z$.

Any system of the form (1) gives rise to a trellis with time-j branches

$$B_j = \{(\mathbf{s}(j), [\mathbf{u}(j), \mathbf{y}(j)], \mathbf{s}(j+1))\}\,, \tag{4}$$

where the branches are labelled with input-output pairs $[\mathbf{u}(j), \mathbf{y}(j)]$; this trellis will be referred to as the *input-output trellis* of the system. Alternatively, we can label the branches with the output symbols $\mathbf{y}(j)$ only, in which case we will refer to the *output trellis* of the system.

The natural definition of observability for a trellis is the following.

Definition [8]: A trellis is *ℓ-observable* if any length-ℓ path segment (sequence of branches) is uniquely determined by the corresponding sequence of branch labels.

This definition is consistent with the traditional system theory notion of observability: the input-output trellis (4) of an ABCD-system (1) is ℓ-observable if and only if the block matrix $[C^T, A^T C^T, \ldots, (A^T)^{\ell-1} C^T]$ has rank m, and it is ℓ-observable for any $\ell \geq m$ if and only if it is m-observable. (The first claim follows from noting that the path segment is uniquely determined by the initial state and the subsequent inputs; the second claim follows from the descending-chain results of [8] or from the Cayley-Hamilton theorem.)

The corresponding matrix condition for the *output* trellis (where the branches are labelled with output symbols $\mathbf{y}(j)$ only) is the following: that trellis is ℓ-observable if and only if the block matrix

$$\begin{bmatrix} C & D & 0 & \ldots & & 0 \\ CA & CB & D & 0 & \ldots & 0 \\ & & \ldots & & & \\ CA^{\ell-1} & CA^{\ell-2}B & \ldots & CAB & CB & D \end{bmatrix} \tag{5}$$

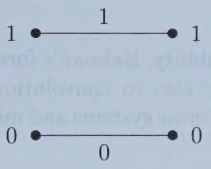


Figure 2: A (section of a) minimal(?) uncontrollable trellis.

has rank $m + \ell k$ (the number of columns); it is ℓ-observable for any $\ell \geq m$ if and only if it is m-observable [1].

4 Controllability Optional

There is also a natural trellis definition of controllability: any state can be reached from any other state in ℓ steps. Other than with observability, its application to ABCD-systems (1) does not dinstinguish between the input-output trellis and the output-only trellis: either trellis is ℓ-controllable if and only if the block matrix $[B, AB, \ldots, A^{\ell-1}B]$ has rank m, and it is ℓ-controllable for any $\ell \geq m$ if and only if it is m-controllable.

On the other hand, the behavioral approach to system theory [9] [10] has revealed that controllability is, in a sense, optional. This is illustrated in Fig. 2: the trellis is minimal for the code that consists of the all-zeros sequence and the all-ones sequence. Rather than ruling out such codes a priori, you may freely choose whether or not to consider bi-infinite sequences through unreachable states as part of the valid behavior. In any case, it is at least required that the trellis is *state-trim*, which means that every state has both a successor and a predecessor (i.e., a bi-infinite path exists through every state); for ABCD-systems (1), the former is automatic and the latter holds if and only if the block matrix $[A, B]$ has rank m.

5 Minimality Simplified

A main result of [8] is a general version of Kalman's theorem (2) for group trellises. Its specialization to linear trellises reads as follows:

Theorem: ([8, Theorem 9], [1, Theorem 5.2]) A time-invariant linear trellis with an m-dimensional state space is minimal if and only if it is state-trim and m-observable.

The theorem applies both to traditional linear systems theory and to convolutional encoders: in the former case, the relevant trellis is the input-output trellis and observability is tested by the rank test of the observability matrix $\mathcal{O}$; in the latter case, the relevant trellis is the output-only trellis and observability is tested by the rank test of the matrix (5).

As discussed in Section 4, you may wish to replace state-trimness by the stronger condition of m-controllability.

6 Conclusion

By a suitable definition of observability, Kalman's formula *"minimal ⇔ controllable and observable"* is now seen to apply also to convolutional encoders. The 35 years long separation between minimality in linear systems and minimality of convolutional encoders is finally overcome.

References

[1] H.-A. Loeliger and T. Mittelholzer, "Convolutional codes over groups," *IEEE Trans. Inform. Theory,* vol. 42, pp. 1660–1686, Nov. 1996.

[2] T. Kailath, *Linear Systems,* Prentice-Hall, 1980.

[3] R. E. Kalman, "On the general theory of control systems," in *Proc. First Int. Congress on Automatic Control,* Butterworth's, London, 1960, pp. 481–493.

[4] J. L. Massey and M. K. Sain, "Inverses of linear sequential circuits," *IEEE Trans. Comput.*, vol. 17, pp. 330–337, April 1968.

[5] G. D. Forney, Jr., "Convolutional codes I: algebraic structure," *IEEE Trans. Inform. Theory,* vol. 16, pp. 720–738, Nov. 1970, and vol. 17, p. 360, May 1971.

[6] G. D. Forney, Jr., "Correction to 'Convolutional Codes I: Algebraic Structure'," *IEEE Trans. Inform. Theory,* vol. 17, pp. 360, May 1971.

[7] G. D. Forney, Jr., R. Johannesson, and Z.-x. Wan, "Minimal and canonical rational generator matrices for convolutional codes," *IEEE Trans. Inform. Theory,* vol. 42, pp. 1865–1880, Nov. 1996.

[8] H.-A. Loeliger, G. D. Forney, Jr., T. Mittelholzer, and M. D. Trott, "Minimality and observability of group systems," *Linear Algebra & Appl.*, vol. 205–206, pp. 937–963, July 1994.

[9] J. C. Willems, "Models for dynamics," in *Dynamics Reported,* vol. 2, U. Kirchgraber and H. O. Walther, Eds., Wiley and Teubner, 1989.

[10] J. C. Willems, "Paradigms and puzzles in the theory of dynamical systems," *IEEE Trans. Aut. Contr.*, vol. 36, pp. 259–294, March 1991.

Trellis-Constrained Codes

Brendan J. Frey
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
`bfrey@turbo.beckman.uiuc.edu`

David J. C. MacKay
Department of Physics, Cavendish Laboratories
Cambridge University
`mackay@mrao.cam.ac.uk`

Abstract

We introduce a class of iteratively decodable *trellis-constrained codes* as a generalization of turbocodes, low-density parity-check codes, serially-concatenated convolutional codes, and product codes. In a trellis-constrained code, multiple trellises interact to define the allowed set of codewords. As a result of these interactions, the minimum-complexity single trellis for the code can have a state space that grows exponentially with block length. However, as with turbocodes and low-density parity-check codes, a decoder can approximate bit-wise maximum *a posteriori* decoding by using the sum-product algorithm on the factor graph that describes the code. We present two new families of codes, *homogenous trellis-constrained codes* and *ring-connected trellis-constrained codes*, and give results that show these codes perform in the same regime as do turbo-codes and low-density parity-check codes.

1 Introduction

Recent interest in the impressive performances of turbocodes and low-density parity-check codes has led to several attempts at generalizing these codes in ways that lead to efficient iterative decoders. In one view that is now quickly propagating across the research network a code is described by graphical constraints on a system of variables, and the iterative decoder is based on a decade-old expert systems algorithm applied to the graph which describes the constraints [1–7]. This *probability propagation algorithm* [8,9] is exact only in cycle-free graphs. However, as evidenced by the excellent error-correcting capabilities of the iterative decoders for turbocodes [10] and low-density parity-check codes [4], the algorithm works impressively well in the graphs that describe these codes, even though they contain many cycles. See [3] and [5] for extensive dissertations on this subject.

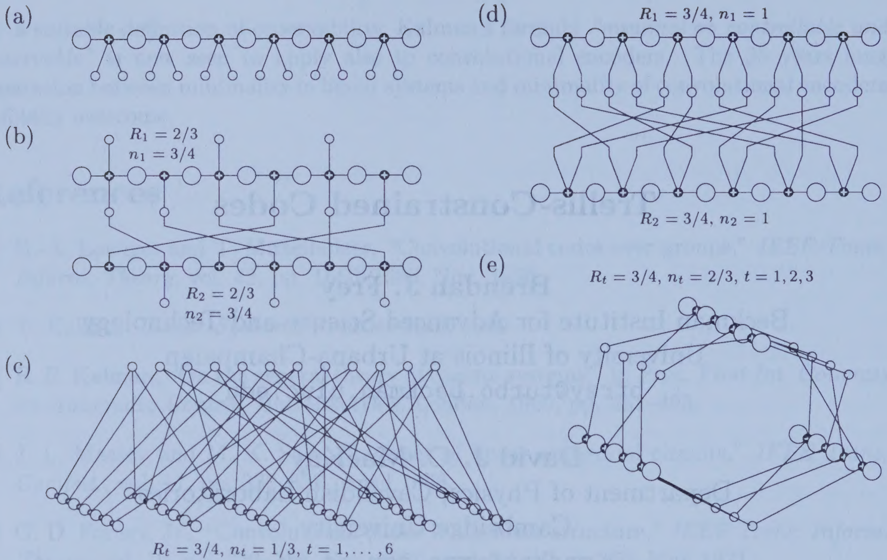


Figure 1: A codeword in a *trellis-constrained code* must simultaneously be a codeword of all the constituent trellises, with the codeword bits reordered. The factor graphs are shown for (a) a single convolutional code; (b) a turbocode, where each parity bit is a codeword bit of only one trellis; (c) a low-density parity-check code; (d) a homogenous trellis-constrained code; and (e) a ring-connected trellis-constrained code. The small unfilled discs represent codeword bits.

Fig. 1a shows the *factor graph* (see [11] in these proceedings) for a trellis. Unlike in a trellis, in a factor graph the values that each state variable (large white discs) can take on are not explicitly shown. Any two adjacent state variables and the corresponding codeword bits (small white discs) must satisfy a linear set of equations (represented by the small black discs with a "+" inside). This representation is called a "factor graph" because it shows how the indicator function for allowed trellis behaviors factors into a product of local functions. Associated with each black disc is a local function of its neighboring variables. Each function evaluates to 1 if its neighboring variables satisfy the local set of linear equations, and to 0 otherwise. The global function is equal to the product of the local functions. A given configuration of the codeword bits is a codeword if the global function evaluates to 1 for some configuration of the state variables. In general, the local functions may be nonlinear, the factor graph variables may be real-valued, and the local functions may evaluate to elements of a semi-ring.

Although factor graphs are less explicit about local relationships than are trellises, factor graphs allow us to represent a richer set of systems. Fig. 1b shows the factor graph for a simple turbocode. A single trellis for the same turbocode would have an unwieldly large number of states. More important than representation, a factor graph provides a framework for iterative decoding via message passing on the graph. The probability propagation algorithm [8, 9], a.k.a. the *sum-product* algorithm [1, 2], can be applied to a factor graph to approximate bit-wise maximum *a posteriori* (MAP) decoding. (In the special case of a turbocode, this general algorithm reduces to turbodecoding [6,7].) Two different factor graphs for the same code may give decoders with different performances.

As another example, Fig. 1c shows the factor graph for a simple low-density parity-check code. Each of the six trellises is a simple parity-check trellis that enforces even parity on its six codeword bits [12].

In a sense, whereas the trellis assisted in the design of low-complexity codes and exact linear-time probabilistic decoders (the Viterbi algorithm and the forward-backward algorithm), the factor graph assists in the design of high-complexity codes and approximate linear-time probabilistic decoders. In this paper, we present a general class of high-complexity, linear-time decodable codes that retain the chain-type structure of trellises locally, as do turbocodes and to a lesser degree low-density parity-check codes. A codeword in a *trellis-constrained code* (TCC) must simultaneously be a codeword of multiple constituent trellises. So, if $f_t(\mathbf{x})$ is the constituent codeword indicator function for trellis $t \in \{1, \ldots, T\}$, the global codeword indicator function is

$$f(\mathbf{x}) = \prod_{i=1}^{T} f_t(\mathbf{x}). \tag{1}$$

Each constituent indicator function is given by a product of the local functions within the corresponding trellis. Usually, the codeword bits interact with the constituent trellises through permuters that rearrange the order of the codeword bits.

For the turbocode in Fig. 1b, there are two constituent functions, $f_1(\cdot)$ and $f_2(\cdot)$. $f_1(\cdot)$ indicates that the upper row of codeword bits are valid output from a convolutional encoder with the middle row of codeword bits as input. $f_1(\cdot)$ does *not* directly place any restrictions on the lower row of codeword bits, so it effectively only checks 3/4 of the codeword bits. $f_2(\cdot)$ indicates that the lower row of codeword bits are valid output from a convolutional encoder with the middle row of codeword bits as input. In contrast to $f_1(\cdot)$, $f_2(\cdot)$ does not place any restrictions on the *upper* row of codeword bits.

The rate R of a TCC is related to the rates of the constituent trellises R_t, $t = 1, \ldots, T$ in a simple way. If n_t is the fraction of codeword bits that trellis t checks, then trellis t removes at most $(1 - R_t)n_tN$ binary degrees of freedom from the code. It may remove a small number less if some of its constraint equations are linearly dependent on those given by other constituent trellises. For large, randomly generated permuters this effect is quite small, so we will ignore it when computing rates in the remainder of this paper. (As a result, the actual rates may be slightly *higher* than the given rates.) The total number of binary degrees of freedom left after all trellis constraints are satisfied is $N - \sum_{t=1}^{T}(1 - R_t)n_tN$, so the rate of the TCC is

$$R = 1 - \sum_{t=1}^{T}(1 - R_t)n_t. \tag{2}$$

From this equation, it is easy to verify that the turbocode in Fig. 1b has rate 1/2 and that the low-density parity-check code in Fig. 1c also has rate 1/2. (Note that a k-bit parity-check trellis has rate $(k - 1)/k$.)

Unlike encoding turbocodes and serially-concatenated convolutional codes, encoding a general TCC takes quadratic time in N. In a general TCC, we can designate a subset of the codeword bits as the systematic bits of the entire code and then use Gaussian elimination to compute a generator matrix (once only). Using such a systematic generator matrix for encoding requires $R(1 - R)N^2$ binary operations.

Decoding a TCC involves performing the forward-backward algorithm for each trellis and exchanging information between trellises in the fashion specified by the sum-product algorithm. The constituent trellises may be processed in parallel or sequentially.

In the following two sections, we present two new families of TCC's and show that they perform in the same regime as do turbocodes and low-density parity-check codes.

2 Homogenous Trellis-Constrained Codes

In a turbocode, the constituent trellises share only a systematic subset of their codeword bits. The other parity bits of each constituent encoder are not constrained by the other encoders. Fig. 1d shows the factor graph for a simple *homogenous TCC* with $T = 2$, in which *all* of the bits are constrained by each constituent trellises. From the general rate formula in (2), we see that the rate for a homegenous turbocode is

$$R = 1 - T(1 - R_{\text{avg}}), \tag{3}$$

where $R_{\text{avg}} = \left(\sum_{t=1}^{T} R_t\right)/T$.

One difference between the homogenous TCC and the turbocode is that the rate of a homogenous TCC decreases directly with the number of trellises T. In the simulations discussed below, we used $T = 2$ and $R_1 = R_2 = R_{\text{avg}} = 3/4$ to get $R = 1/2$. To obtain the same rate with $T = 3$ would require $R_{\text{avg}} = 5/6$. In contrast, the rate for a turbocode varies roughly inversely with T. A rate 1/2 turbocode with $T = 3$ can be obtained with $R_1 = R_2 = R_3 = 3/4$. Another difference is that the permuter length of a homogenous TCC is N, whereas for a turbocode, the permuter length is RN.

2.1 Encoding and Decoding

The trellises in a homogenous TCC share all their bits, so we can't simply encode by dividing the bits in each constituent trellis into a systematic set and a parity set and running a linear-time encoding method for each trellis, as is possible in a turbocode. Instead, we apply a previously computed generator matrix to a previously selected systematic subset of codeword bits, which takes $R(1 - R)N^2$ binary operations.

The iterative decoder processes each constituent trellis using the forward-backward algorithm, and passes "extrinsic information" between the trellises in the manner specified by the sum-product algorithm. For two trellises, the decoding schedule is straightforward. For $T > 2$, different decoding schedules are possible. The trellises may be processed sequentially, in which case the current trellis uses the most recently computed probabilities produced by the other trellises. Alternatively, the trellises may be processed in parallel, in which case the current trellis uses the probabilities produced by the other trellises in the previous decoding iteration.

For the sake of gaining insight into these new compound codes and the behavior of their iterative decoders, we prefer to decode until a codeword is found or we are sure the iterative decoder has failed to find a codeword. After each decoding iteration, the current bit-wise MAP estimates are used to determine whether a valid codeword has been found, in which case the iterative procedure is terminated. If 100 iterations are completed without finding a codeword, we label the block a decoding failure. Notice that given the factor graph of a code, determining that a codeword is valid is simply a matter of checking that all the local functions evaluate to 1.

2.2 Performance on an AWGN Channel

Using Monte Carlo, we estimated the performance of an $N = 131,072$, $T = 2$ homogenous TCC with $R_1 = R_2 = 3/4$, giving $R = 1/2$. (See App. A for a description of how the BER confidence intervals were computed.) Each rate 3/4 trellis was obtained by shortening every fifth bit of a rate 4/5 nonsystematic convolutional code with maximum d_{min}. (The generator polynomials for this code are given in [13] and are $(32, 4, 22, 15, 17)_{\text{octal}}$.) Fig. 2 shows the performance of this homogenous TCC, relative to the turbocode introduced by Berrou *et. al.* [14] and the best rate 1/2, $N = 65,389$ low-density parity-check code published to date [4]. Although it does not perform as well as the turbocode, it performs significantly better than the low-density parity-check code. We believe there is room for improvement here, since we chose the set of generator polynomials that gave maximum d_{min} and this is quite likely not the best choice. (Keep in mind, however, that the performance of a homogenous TCC is not necessarily governed by the same constituent trellis properties that govern the performance of a turbocode.) Of significant importance compared to turbocodes, we have observed that this homogenous TCC does not have low-weight codewords.

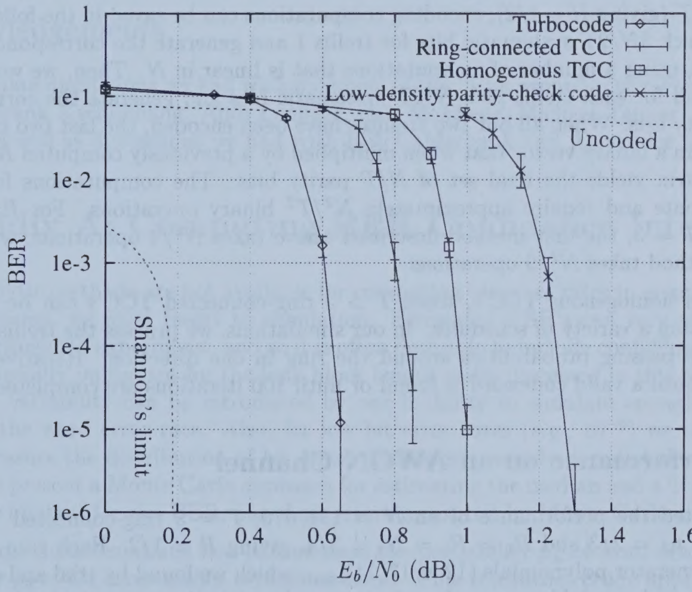


Figure 2: The performances of a homogenous TCC and a ring-connected TCC compared to the best rate 1/2 turbocode and low-density parity-check code performances published to date [4, 10].

3 Ring-Connected Trellis-Constrained Codes

Fig. 1e shows the factor graph for a simple *ring-connected TCC* with $T = 3$. This code can be viewed as a serially-concatenated convolutional code [15, 16] in which some of the output bits are constrained to be equal to some of the input bits. The factor graph thus forms a ring of connected trellises. In the ring-connected TCC shown, each constituent trellis checks exactly $2/T$ of the codeword bits. From the general rate formula in (2), the rate for such a ring-connected turbocode is

$$R = 2(R_{\text{avg}} - 1/2). \tag{4}$$

Unlike turbocodes and homogenous TCC's, for such a ring-connected TCC the rate does not depend on the number of constituent trellises T. However, the permuter lengths are $1/T$ relative to the block length.

3.1 Encoding and Decoding

For ring-connected TCC's, we cannot use the same type of encoding procedure that is used for serially-concatenated convolutional codes, since some of the "output" bits must match some of the "input" bits. As with the homogenous TCC, we can encode with approximately $R(1-R)N^2$ binary operations by applying a previously computed generator matrix to a previously selected systematic subset of codeword bits. However, if $R_t = 3/4$, $t = 1, \ldots, T$ (giving $R = 1/2$), encoding computations can be saved in the following way. First, we pick $3N/2T$ systematic bits for trellis 1 and generate the corresponding $N/2T$ parity bits, using a number of computations that is linear in N. Then, we work around the ring and for each trellis pick $N/2T$ systematic bits and generate the corresponding $N/2T$ parity bits. When all but two trellises have been encoded, the last two trellises are used to form a binary vector that when multiplied by a previously computed $N/T \times N/T$ binary matrix yields the final set of N/T parity bits. The computations for the last step dominate and require approximately N^2/T^2 binary operations. For $R_{\text{avg}} = 3/4$, $R = 1/2$, $T = 3$, the first method described above takes $N^2/4$ operations, whereas the second method takes $N^2/9$ operations.

As with homogenous TCC's, when $T > 2$ ring-connected TCC's can be iteratively decoded using a variety of schedules. In our simulations, we process the trellises sequentially while passing probabilities around the ring in one direction. Iterative decoding continues until a valid codeword is found or until 100 iterations are completed.

3.2 Performance on an AWGN Channel

We estimated the performance of an $N = 131,070$, $T = 3$ ring-connected TCC with $n_1 = n_2 = n_3 = 2/3$ and $R_1 = R_2 = R_3 = 3/4$, giving $R = 1/2$. Each rate 3/4 trellis used the generator polynomials $(12, 4, 17, 11)_{\text{octal}}$, which we found by trial and error. The constituent codeword bits were shared alternately with the two neighboring trellises. Fig. 2 shows the performance of this ring-connected TCC. It performs signficantly better than the homogenous TCC and the low-density parity-check code and only 0.2 dB worse than the turbocde.

4 Discussion

There is some literature on the properties of the constituent trellises that make good turbocodes [10, 17, 18] and serially-concatenated convolutional codes [19]. We are currently exploring similar arguments for choosing the properties of the constituent trellises that will make good homogenous TCC's and ring-connected TCC's. The behavior of these two new families of TCC's is quite different from that of turbocodes, so we expect different properties to be relevant. For example, in order to avoid low-weight codewords in a turbocode, we try to avoid permuters that keep a pair of bits the same distance apart in the two constituent trellises [10]. This degenerate effect is "broken" by the ring of a ring-connected TCC, which requires not only that two neighboring trellises have an equal set of shared "input" bits, but also that their "output" bits must satisfy the constraints given by the ramainder of the ring. However, in a ring-connected TCC low-weight codewords are introduced by permuters that preserve a degenerate pattern around the entire ring. Another important question is what properties of a TCC lead to the initial drop (say down to a BER of 0.01)? These properties may very well be independent (or even contrary to) those properties that give high minimum distance.

We believe that the general class of "trellis-constrained codes" presented in this paper is a fruitful generalization of several other iteratively decodable codes. If we think of entire trellises as nodes in a graph whose edges represent bundles of codeword bits, it is evident that there are a variety of other interesting new TCC's aside from those shown in Fig. 1 that have yet to be explored.

Acknowledgements

We appreciate discussions we had with members of the "Ulm group", Dave Forney, Ralph Koetter, Frank Kschischang, Andi Loeliger, Bob McEliece, Michael Tanner, and Niclas Wiberg, as well as discussions we had with Rob Calderbank and Alexander Vardy.

Appendix A: Computing BER Confidence Intervals

When analytic methods are not available for computing bit error rates in error-correcting coding systems, we must resort to simulation. Estimated BER's can vary significantly from experiment to experiment, and so it is often desirable to include confidence intervals. This is especially important for the long block length codes discussed in this paper, since significant variability can be introduced by our inability to simulate enough blocks to pin down the word error rate. Also, for low bit error rates (*e.g.*, 10^{-6}) we may not be able to measure the distribution of bit errors within erroneously decoded words. In this section, we present a Monte Carlo approach for estimating the median and a 2.5% / 97.5% confidence interval for the BER, based on errors measured by Monte Carlo simulation.

The error model contains two parameters: the probability p_w of word error, and the probability p_b of bit error *within erroneous words.* This is a rather crude approximation, since in practice we expect there to be more than one failure mode, *i.e.*, there ought to be several p_b's corresponding to different failure modes.

Let M be the number of words transmitted and let n_w be the number of measured word errors. Let K be the number of information bits per word, and let n_b be the *total* number of bit errors measured while transmitting all M blocks. We will refer to the measured values as the data, $\mathcal{D} = \{n_w, n_b\}$. From the Bayesian perspective, before observing $\mathcal{D}$, we place a prior distribution $p(p_w, p_b)$ on the error model parameters. After observing $\mathcal{D}$, we draw conclusions (*e.g.*, compute a confidence interval) from the posterior distribution $p(p_w, p_b|\mathcal{D})$, where

$$p(p_w, p_b|\mathcal{D}) \propto p(p_w, p_b)P(\mathcal{D}|p_w, p_b). \qquad (5)$$

In this equation, the constant of proportionality does not depend on p_w or p_b. The last factor $P(\mathcal{D}|p_w, p_b)$ is called the likelihood.

We let p_w and p_b be independent beta-distributed random variables under the prior: $p(p_w, p_b) = p(p_w)p(p_b)$, where

$$p(p_w) \propto p_w^{\alpha_w - 1}(1 - p_w)^{\beta_w - 1}, \quad \text{and} \quad p(p_b) \propto p_b^{\alpha_b - 1}(1 - p_b)^{\beta_b - 1}. \tag{6}$$

In frequentist terms, α_w and β_w have the effect of shrinking our measurements toward a word error rate of $\alpha_w/(\alpha_w + \beta_w)$, where the influence of this shrinkage grows with $\alpha_w + \beta_w$. Typically, we choose $\alpha_w = \beta_w = 1$, which gives a uniform prior over p_w as shown in Fig. 3a.

As for the prior over p_b, it should be chosen while keeping in mind the behavior of the decoder. If the main source of bit errors is a failure to decode, and if we believe that for failures the decoder will produce a probability of bit error that is roughly equal to the probability p_u of bit error for uncoded transmission, then the prior should place weight on $p_b = p_u$. In this case, we choose $\alpha_b = 2$ and $\beta_b = 1/p_u$, which ensures that the mode of the prior occurs at p_u and that the prior is relatively broad. For example, for $E_b/N_0 = 1$ dB we have $p_u = 0.0563$, and so We choose $\alpha_b = 2$ and $\beta_b = 1/0.0563 = 17.76$, giving the prior distribution for p_b shown in Fig. 3b.

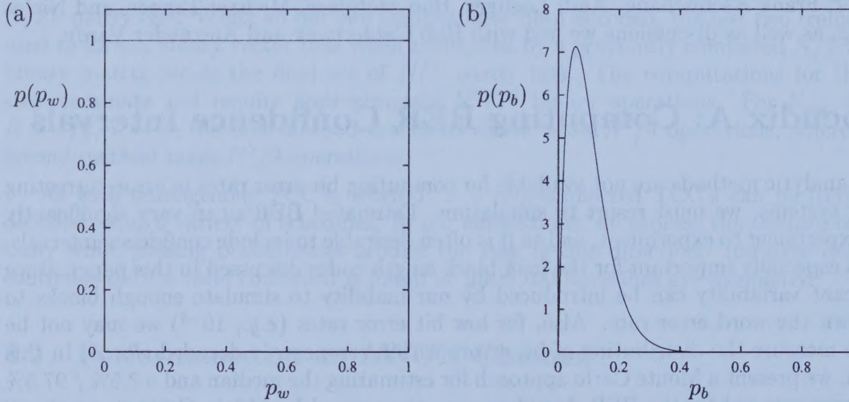


Figure 3: (a) The prior distribution over the probability of word error p_w. (b) The prior distribution over the probability of bit error p_b within erroneous words. This distribution is designed so that its median is equal to the probability of bit error for uncoded transmission.

It is straightforward to show that the likelihood is

$$P(\mathcal{D}|p_w, p_b) = P(n_w, n_b|p_w, p_b) \propto p_w^{n_w}(1 - p_w)^{M - n_w} p_b^{n_b}(1 - p_b)^{n_w K - n_b}. \tag{7}$$

This distribution is the product of a binomial distribution for the number of word errors and a binomial distribution for the number of bit errors. Combining this likelihood with the prior, we obtain the posterior,

$$p(p_w, p_b|\mathcal{D}) \propto p_w^{\alpha_w - 1 + n_w}(1 - p_w)^{\beta_w - 1 + M - n_w} p_b^{\alpha_b - 1 + n_b}(1 - p_b)^{\beta_b - 1 + n_w K - n_b}, \tag{8}$$

which is just the product of a beta distribution over p_w and a separate beta distribution over p_b. Of course, we are actually interested in the posterior distribution $p(p_w p_b|\mathcal{D})$ over the total probability of a bit error $p_w p_b$. A sample is obtained from $p(p_w p_b|\mathcal{D})$ by drawing $p_w - p_b$ pairs from the posterior in (8) and taking the product of p_w and p_b in each pair. This sample is sorted in ascending order, and the value of $p_w p_b$ occuring half-way through the sorted list is taken as an estimate of the median of $p(p_w p_b|\mathcal{D})$. Similarly, the values of $p_w p_b$ occuring 2.5% and 97.5% through the sorted list are taken as the 95% confidence interval.

For the homogenous TCC, we simulated the transmission of $M = 332$ blocks at $E_b/N_0 = 0.95$ dB using a block length of $N = 131,072$. We measured $n_w = 14$ and $n_b = 34,225$. Using the prior presented above for the slightly higher value of $E_b/N_0 = 1$ dB, a sample of 1000 points from the posterior over p_w and p_b was obtained and is shown in Fig. 4a. As described above, for $\gamma = 0.025$, 0.5 and 0.975, we found the values for p_γ such that $\hat{p}(p_w p_b < p_\gamma|\mathcal{D}) = \gamma$, where $\hat{p}$ is the sample distribution. The corresponding three curves of the form $p_w p_b = p_\gamma$ are shown in Fig. 4a, and the corresponding values of p_γ give a median of 1.7×10^{-3} and a 95% confidence interval of $(9.9 \times 10^{-4}, 2.6 \times 10^{-3})$. Clearly, in this case it is the values for p_w that determine the p_γ's for these curves, whereas the values for p_b are well-determined by the measurements. We could have assumed that p_b took on its measured value instead of sampling from the posterior.

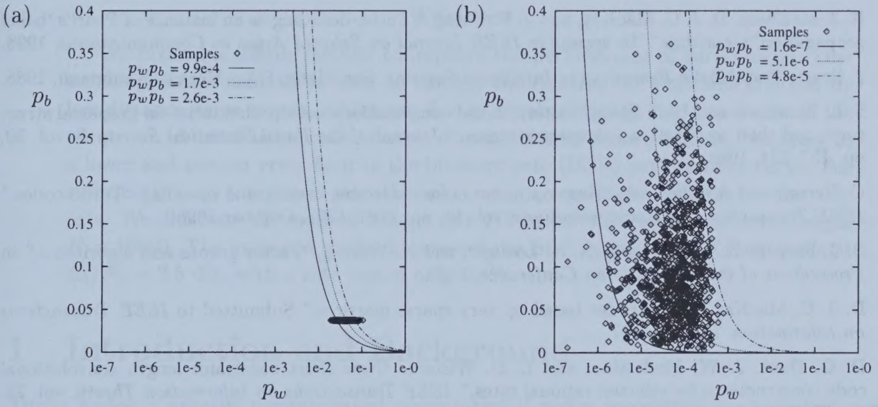


Figure 4: (a) A 1000-point sample from $p(p_w, p_b|\mathcal{D})$ for $M = 332$, $n_w = 14$, $K = 65,536$ and $n_b = 34,225$, for the prior described in the main text. (b) A 1000-point sample from $p(p_w, p_b|\mathcal{D})$ for $M = 10,216$, $n_w = 0$, $K = 65,536$ and $n_b = 0$, for the same prior.

For the homogenous TCC described above, we also simulated the transmission of $M = 10,216$ blocks at $E_b/N_0 = 1.0$ dB. In this case, we measured $n_w = 0$ and $n_b = 0$. Using naive methods, we might conclude that the bit error rate is 0 and that there isn't any variation in this value. However, the Bayesian technique gives the sample from the posterior shown in Fig. 4b. In this case, the values of both p_w and p_b play a role in determining the p_γ's for the three curves. The median is 5.1×10^{-6} and the confidence interval is $(1.6 \times 10^{-7}, 4.8 \times 10^{-5})$.

References

[1] R. M. Tanner, "A recursive approach to low complexity codes," *IEEE Transactions on Information Theory*, vol. 27, pp. 533–547, 1981.

[2] N. Wiberg, H.-A. Loeliger, and R. Kötter, "Codes and iterative decoding on general graphs," *European Transactions on Telecommunications*, vol. 6, pp. 513–525, September/October 1995.

[3] N. Wiberg, *Codes and Decoding on General Graphs.* Linköping Sweden: Department of Electrical Engineering, Linköping University, 1996. Doctoral dissertation.

[4] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *Electronics Letters*, vol. 32, pp. 1645–1646, August 1996. Due to editing errors, reprinted in *Electronics Letters*, vol. 33, March 1997, 457–458.

[5] B. J. Frey, *Bayesian Networks for Pattern Classification, Data Compression and Channel Coding.* Toronto Canada: Department of Electrical and Computer Engineering, University of Toronto, 1997. Doctoral dissertation available at `http://www.cs.utoronto.ca/~frey`.

[6] F. R. Kschischang and B. J. Frey, "Iterative decoding of compound codes by probability propagation in graphical models." To appear in *IEEE Journal on Selected Areas in Communications*, available at `http://www.cs.utoronto.ca/~frey`, 1998.

[7] R. J. McEliece, D. J. C. MacKay, and J. F. Cheng, "Turbo-decoding as an instance of Pearl's 'belief propagation' algorithm." To appear in *IEEE Journal on Selected Areas in Communications*, 1998.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems.* San Mateo CA.: Morgan Kaufmann, 1988.

[9] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society B*, vol. 50, pp. 157–224, 1988.

[10] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Transactions on Communications*, vol. 44, pp. 1261–1271, October 1996.

[11] B. J. Frey, F. R. Kschischang, H. A. Loeliger, and N. Wiberg, "Factor graphs and algorithms," in *Proceedings of the* 35^{th} *Allerton Conference*, 1998.

[12] D. J. C. MacKay, "Good codes based on very sparse matrices." Submitted to *IEEE Transactions on Information Theory*, 1997.

[13] D. G. Daut, J. W. Modestino, and L. D. Wismer, "New short constraint length convolutional code constructions for selected rational rates," *IEEE Transactions on Information Theory*, vol. 28, pp. 794–800, September 1982.

[14] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," in *Proceedings of the IEEE International Conference on Communications*, 1993.

[15] S. Benedetto and G. Montorsi, "Serial concatenation of block and convolutional codes," *Electronics Letters*, vol. 32, pp. 887–888, 1996.

[16] S. Benedetto and G. Montorsi, "Iterative decoding of serially concatenated convolutional codes," *Electronics Letters*, vol. 32, pp. 1186–1188, 1996.

[17] S. Benedetto and G. Montorsi, "Unveiling turbo-codes: Some results on parallel concatenated coding schemes," *IEEE Transactions on Information Theory*, vol. 42, pp. 409–428, March 1996.

[18] D. Divsalar and R. J. McEliece, "Effective free distance of turbocodes," *Electronics Letters*, vol. 32, pp. 445–446, February 1996.

[19] S. Benedetto, G. Montorsi, D. Divsalar, and F. Pollara, "Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding." To appear in *IEEE Transactions on Information Theory*, 1997.

Application of Distance Spectrum Analysis to Turbo Code Performance Improvement

Mats Öberg* and Paul H. Siegel
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, California 92093-0407
E-mail: moberg@ucsd.edu, psiegel@ucsd.edu

Abstract

We present a simple method to improve the performance of turbo codes. Distance spectrum analysis is used to identify information bit positions affected by low-distance error events, which are few in number due to the sparseness of the spectrum. A modified encoder inserts dummy bits in these positions, resulting in a lower and steeper error floor in the bit-error-rate (BER) performance curve. For sufficiently large interleaver size, the only cost is a very slight reduction in the code rate. We illustrate the method using a rate 1/2 turbo code, with interleaver length $N = 10000$. The proposed method improves the BER by an order of magnitude at $E_b/N_0 = 2.5$ dB, with a rate loss of only 0.4%.

1 Introduction and Background

Turbo codes – parallel concatenated convolutional codes connected by an interleaver – were first introduced in 1993 by Berrou et. al [1], and are now widely recognized as a landmark development in error control coding. Two characteristic features of turbo code performance are the small bit-error-rate (BER) achieved even at very low signal-to-noise ratio (SNR) and the flattening of the error-rate curve – the so-called "error floor" – at moderate and high values of SNR.

Recently Perez et. al. [2] analyzed the codeword weight distribution and multiplicity – the distance spectrum – of turbo codes and offered an explanation for both of these phenomena. By considering input sequences of length equal to the interleaver size N, they derived a maximum-likelihood (ML) bound for the BER performance of turbo codes,

$$P_b \leq \sum_{d=d_{free}}^{2(\nu+N)} \frac{N_d \tilde{\omega}_d}{N} Q\left(\sqrt{d\frac{2RE_b}{N_0}}\right) \tag{1}$$

where N_d is the multiplicity of weight-d codewords, $\tilde{\omega}_d$ is the average weight of the information sequences causing weight-d codewords, R is the code rate, and ν is the

---

[*] This research is supported by the Royal Swedish Academy of Sciences and the Center for Wireless Communications at UCSD. The material in this paper was presented in part at the International Symposium on Turbo Codes & Related Topics, Brest, France, September 3-5, 1997.

memory of the constituent convolutional code. Although the iterative decoding procedure utilized in turbo codes is sub-optimal [3], its performance has been found to be very close to that of an optimal ML decoder. Thus the bound in (1) can be used to evaluate turbo code performance.

At moderate to high SNR, the BER bound is dominated by the free-distance term, and the performance approaches the *free-distance asymptote*, P_{free}, given by

$$P_{free} = \frac{N_{free}\tilde{\omega}_{free}}{N} Q\left(\sqrt{d_{free}\frac{2RE_b}{N_0}}\right), \tag{2}$$

where d_{free} denotes the free distance of the code, with corresponding multiplicity N_{free} and average information weight $\tilde{\omega}_{free}$.

In [2], Perez, *et al.* studied a turbo code based upon a particular pseudo-random interleaver with size $N = 65536$. The code was found to have $d_{free} = 6$, $N_{free} = 3$, and $\tilde{w}_{free} = 2$. The simulated BER performance was shown to approach the free-distance asymptote, which had a shallow slope as a result of the relatively small free-distance of the turbo code. The observed error floor could therefore be interpreted as simply a manifestation of the relative flatness of the asymptote.

For this particular turbo code, the free distance codewords all have information weight 2, as well as a very small effective multiplicity,

$$\frac{N_{free}}{N} = \frac{3}{65536},$$

resulting in a small error coefficient in the expression (2) for the free-distance asymptote P_{free}. The small error coefficient accounts for the fact that, despite the small free-distance of this turbo code, the error floor appears at a rather low error rate ($BER \approx 10^{-6}$ at SNR=1 dB).

The analysis of "average" turbo codes in [2] shows that this is the typical situation and, moreover, the unusually small multiplicity extends to low-weight codewords in general. This "spectral sparseness" [2], which is not found in conventional convolutional codes, can be attributed to the parallel concatenated convolutional code structure of turbo codes and the properties of pseudo-random interleavers. A consequence of this spectral sparseness is that the free-distance asymptote dominates the error-rate bound even at very low SNR. This fact, along with the small effective multiplicity for free-distance codewords, accounts for the remarkable performance of turbo codes in the SNR region close to the Shannon limit.

Based upon this analysis of turbo code distance spectrum properties, two modifications to the turbo code design that would improve overall code performance were suggested in [2]. The error floor can be lowered, without a slope change, by increasing the interleaver size N, while maintaining the free distance and corresponding multiplicity. Alternatively, the slope of the error floor can be increased (negatively) by choosing constituent convolutional codes that increase the free distance of the turbo code, while keeping the corresponding multiplicity small.

In this paper, we will propose another approach to modifying the turbo code that does not require any change of the interleaver size or the constituent convolutional codes. By examining the interplay between the constituent convolutional codes and the pseudo-random interleaver, we derive a list of the non-zero bit positions in the information frames that produce minimum-weight turbo codewords. A modified encoder places dummy bits in these positions, and after the turbo decoding is completed, the contents of these positions in the decoded frame are discarded. This process effectively removes the contribution to the BER of the free-distance error events, resulting in a lowering of the error floor and a change in its slope. By determining bit positions that correspond to other low-weight codewords, we can further improve code performance by applying this same procedure to those frame positions. The small multiplicity of low-weight codewords in the turbo code implies that the rate loss incurred by this encoder modification is negligible for large interleaver size.

The remainder of the paper is organized as follows. In Section 2, we discuss properties of the constituent convolutional encoders that play a role in the analysis of turbo code performance. In Section 3, we investigate the effect of the interleaver on the distance spectrum and, in particular, the set of low-weight codewords. Section 4 provides the details of the new method for improving the turbo code performance, as well as BER simulation results that confirm the expected gains. Section 5 summarizes our results.

2 Properties of the Constituent Encoders

A turbo encoder consists of two or more recursive systematic convolutional (RSC) encoders in parallel concatenation, along with interleavers which permute the input bits of the first encoder before applying them to the inputs of the other encoders. Input bits in a frame of length N are encoded by the first RSC, producing what we call the first dimension codeword. The same information frame is permuted by the interleaver and encoded by the second encoder, generating the second dimension sequence. A similar procedure is followed with any additional encoders. Since the constituent encoders are systematic, only the first information frame need be transmitted, along with the parity bits from each encoder. In this paper, we will consider only the case of two identical, rate 1/2 RSC encoders. To increase the overall rate of the encoder from 1/3 to 1/2, we follow the usual practice of puncturing every other parity bit in each dimension.

Berrou, *et al.* [4] proved a property of RSC encoders that plays an important role in the characterization of minimum-distance error events in turbo encoders. They showed that if $h(D)$ is the feedback polynomial of a RSC encoder, and if $h(D)$ is a divisor of $1 + D^L$, then $h(D)$ is also a divisor of $1 + D^{pL}$ for any integer $p \geq 1$. Let 0^i denote a run of 0's of length i. The result implies that an input stream $1\,0^{Lp-1}\,1$, where the two ones are separated by $pL - 1$ zeros, for any integer $p \geq 1$, will generate a trellis path that diverges from the all-zero path and then remerges after $pL + 1$ trellis steps. The encoder will leave the all-zero state in response to the first 1, and then will return to that state in response to the second 1.

Example 1 *Consider the memory-4, RSC encoder with parity-check polynomials* $h_0(D) = 1 + D + D^2 + D^3 + D^4$ *(octal 37) and* $h_1(D) = 1 + D^4$ *(octal 21) [2]. The feedback polynomial* $h_0(D)$ *divides* $1 + D^5$*. Therefore, any binary sequence* $1\,0^{5p-1}\,1$, $p \geq 1$, *will force the encoder to leave the all-zero state at the appearance of the first 1, and return to that state* $5p + 1$ *steps later, in response to the second 1. Fig. 1 illustrates the codeword corresponding to the case* $p = 1$ *. Note that the weight of the resulting codeword is 6.*

Example 2 *For an encoder with the same feedback polynomial* $h_0(D)$ *as in Example 1, any binary sequence* $h_0(D)^r$, $r \geq 1$, *will force the encoder to leave the all-zero state at the first 1, and return at the final 1. When* $r = 1$, *one gets the length-5 sequence* $1\,1\,1\,1\,1$, *which generates a weight-7 codeword.*

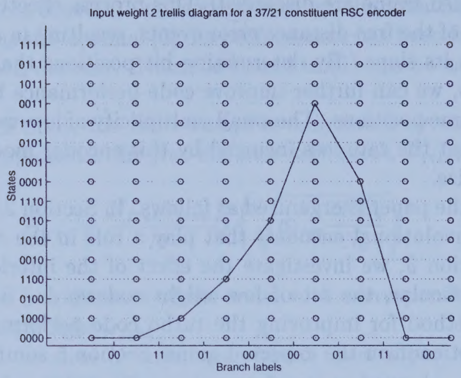


Figure 1: Response of a octal (37,21) RSC encoder with input pattern 0100001000.

The linearity of RSC encoders permits the interpretation of these observations in terms of error events. In the case of the feedback polynomial in the preceding examples, any pair of code sequences $a(D)$ and $b(D)$ that differ from another by $e(D) = 1 + D^{5p}$, $p \geq 1$, or $e(D) = h_0(D)^r$, $r \geq 1$, will correspond to trellis paths whose state sequences are distinct during the course of an error event of length equal to the degree of $e(D)$.

For a RSC encoder, one might expect that low-weight code sequences would be generated in response to low-weight input sequences that correspond to short error events in the trellis. In particular, the weight-2 input sequences discussed in Example 1, in which the non-zero bits are separated by a small multiple of the code constraint length $L = 5$, are good candidates for producing minimum-weight code sequences. For the rate 1/2, octal (37,21) code, the free distance is $d_{free} = 6$, and Fig. 1 depicts a minimum distance code sequence that is generated in response to a weight-2 input sequence in which the non-zero bit separation is one constraint length.

As discussed in [2], for the constituent convolutional code, the multiplicity of free-distance error events, N_{free}, will be roughly proportional to the frame length N, since there are few restrictions on the frame positions where the event may begin. In the next section, we will see that the introduction of an interleaver in a parallel concatenated RSC coding structure will dramatically restrict the possible starting positions of free-distance error events.

3 The Effect of the Interleaver

The remarkable power of turbo codes comes from the combination of the parallel concatenated RSC codes and the permutation of the information frame by the interleaver. The interleaver permutes the frame of information bits in the first dimension prior to their encoding by the encoder in the second dimension. Loosely speaking, the effect of this permutation is to ensure that low-weight error events occur in only one dimension. Depending on the choice of the permutation, the interleaver can affect both the distances and the multiplicities of error events.

Low-weight turbo codewords arise when the interleaver maps low-weight information frames that produce low-weight parity in the first dimension into frames that produce low-weight parity in the second dimension. In view of the discussion in the previous section, therefore, the interleaver should avoid certain permutations of bit positions as much as possible. Specifically, define the polynomial $b(D)$ by

$$b(D) = \sum_{n=0}^{p\nu} b_n D^n = h_0(D)^p, \tag{3}$$

and let $\mathcal{M}_p$ denote the indices of non-zero coefficients of $b(D)$,

$$\mathcal{M}_p = \{n | b_n \neq 0\}. \tag{4}$$

Then, the interleaver should avoid the following mappings of subsets of bit positions:

$$\begin{aligned} \{i, i + Lp\} &\longmapsto \{j, j + Lr\} & (5) \\ \{i, i + Lp, j, j + Lr\} &\longmapsto \{k, k + Lu, m, m + Lv\} & (6) \\ \{i + \mathcal{M}_p\} &\longmapsto \{j + \mathcal{M}_r\} & (7) \end{aligned}$$

where L is the parameter in the divisibility condition of the previous section; i, j, k and m are positive integers $\leq N$; ν is the memory length of the constituent RSC encoder; and r, s, u and v are integers small enough that the corresponding information sequences produce low parity weight at the encoder output. Note that, in the mappings in (7) above, we require that $|\mathcal{M}_p| = |\mathcal{M}_r|$.

Remark: Permutations that map unions of the various subsets above to other unions of these subsets having equal information weight should also be avoided.

Analysis of random interleavers shows that an "average interleaver" can accomplish these desired objectives fairly well [5], and pseudo-randomly generated interleavers have been found to be generally superior to other structured interleavers. The superior performance does not typically arise from a large free distance, however. Rather, the best performing interleavers appear to gain their advantage through drastic reduction of the low-distance error event multiplicity. This "spectral thinning" is the property demonstrated through example and analysis in [2], and lies at the heart of the technique for improving turbo code performance as described in the next section.

We analyzed the distance spectrum of a turbo code based upon the aforementioned rate 1/2 constituent code and a particular pseudo-random interleaver of size $N = 10000$. (In the analysis, we assume that, at the end of the information frame, the encoders in both dimensions are driven to the all-zero state by appending appropriate tails of length ν.) The free distance of the turbo code is $d_{free} = 6$, with free-distance multiplicity $N_{free} = 4$. As was the case for the $N = 65536$ turbo code discussed in [2], all of the free-distance codewords correspond to weight-2 information sequences. We also determined the codewords of weight 16 or less that are generated by weight-2 inputs where the non-zero information bit positions in both dimensions are separated by a small multiple of the constituent code constraint length. In our search for mappings of bit patterns corresponding to powers of the feedback polynomial that mapped into a bit pattern corresponding to a power of the feedback polynomial, we found none which would result in codeword weight 16 or less.

Table 1 describes in more detail the search results for codewords of weight 10 or less. The table specifies the positions of the non-zero bits in information frames of weight 2 that generate code sequences with weight 6, 8, or 10. For each bit pair, the bit separation in the first dimension and the bit separation in the second dimension after permutation by the interleaver are indicated in the column labeled "Mapping". The bit separation is

Mapping	Bit pairs for $d = 6$ (dimension 1)	$(w_H(P_1), w_H(P_2))$
$1 \mapsto 1$	$(1979, 1984)(5861, 5866)$	$(2,2)(2,2)$
$1 \mapsto 2$	$(785, 790)(1383, 1388)$	$(2,2)(2,2)$

Mapping	Bit pairs for $d = 8$ (dimension 1)	$(w_H(P_1), w_H(P_2))$
$1 \mapsto 2$	$(5117, 5122)(5827, 5832)(9014, 9019)$	$(2,4)(2,4)(2,4)$
$1 \mapsto 3$	$(449, 454)(1280, 1285)(5566, 5571)$	$(2,4)(2,4)(2,4)$
$2 \mapsto 1$	$(4253, 4263)$	$(4,2)$
$2 \mapsto 2$	$(1163, 1173)$	$(4,2)$
$3 \mapsto 1$	$(3518, 3533)(7132, 7147)$	$(4,2)(4,2)$
$3 \mapsto 2$	$(5804, 5819)(9136, 9151)$	$(4,2)(4,2)$
$4 \mapsto 2$	$(3132, 3152)$	$(4,2)$

Mapping	Bit pairs for $d = 10$ (dimension 1)	$(w_H(P_1), w_H(P_2))$
$1 \mapsto 6$	$(1982, 1987)$	$(2,6)$
$2 \mapsto 2$	$(3785, 3795)$	$(4,4)$
$2 \mapsto 4$	$(3026, 3036)$	$(2,6)$
$3 \mapsto 2$	$(2368, 2383)(5265, 5280)(8061, 8076)$	$(4,4)(4,4)(4,4)$
$3 \mapsto 3$	$(2528, 2543)(4925, 4940)(9944, 9959)$	$(4,4)(4,4)(4,4)$
$3 \mapsto 4$	$(2040, 2055)$	$(4,4)$
$4 \mapsto 1$	$(4053, 4073)(6393, 6413)$	$(6,2)(6,2)$
$4 \mapsto 2$	$(4823, 4843)(9406, 9426)$	$(6,2)(4,4)$
$5 \mapsto 1$	$(4398, 4423)(7542, 7567)(8793, 8818)(9000, 9025)$	$(6,2)(6,2)(6,2)(6,2)$
$6 \mapsto 1$	$(7648, 7678)$	$(6,2)$

Table 1: Weight-2 inputs generating turbo codewords of weights 6, 8, and 10.

given in terms of the number of constraint lengths $L = 5$ that it spans. The table also shows the weight of the parity bits in the first and second dimensions. The total number of non-zero information bit positions in the calculated codewords of weights 6, 8, 10, 12, 14, and 16 are 8, 26, 38, 34, 24, and 8, respectively.

For weight-6 codewords, the list is complete and our simulation results indicate that that is the case also for weight-8 and weight-10 codewords. Note that these free-distance codewords correspond to input frames of weight 2 where the bit separation in the first dimension is one constraint length, and the corresponding bit separation in the second dimension is either one or two constraint lengths.

4 Improving BER by Discarding Spectral Lines

Fig. 2 shows the simulated BER as a function of SNR for the $N = 10000$ turbo code of the previous section. Also shown are the free-distance asymptote and the weight-8 asymptote corresponding to the codewords listed in Table 1.

Fig. 3 shows a histogram of the cumulative number of errors that occurred at each bit position in the information frame for SNR=2 dB. A total of 2,177,440,000 information bits were encoded and 3501 bits were decoded in error, yielding a BER of 1.61×10^{-6}. Of these bit errors, 2907 occurr at positions associated with the codewords listed in Table 1, and another 52 errors are associated with $d = 12$, $d = 14$ and $d = 16$ codewords found in our search. Specifically, 1676 bit errors correspond to $d = 6$ codewords, 952 to $d = 8$, 279 to $d = 10$, 44 to $d = 12$, 8 to $d = 14$, and 0 to $d = 16$.

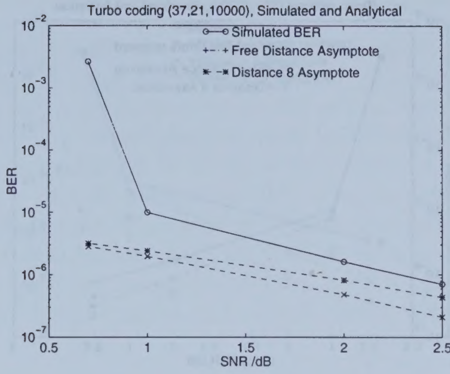


Figure 2: Simulated BER and asymptotes for distances 6 and 8.

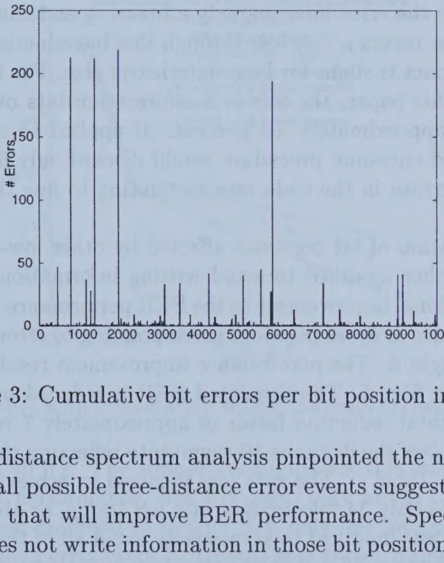


Figure 3: Cumulative bit errors per bit position in frame.

The fact that the distance spectrum analysis pinpointed the non-zero information bit positions involved in all possible free-distance error events suggests a simple modification to the turbo encoder that will improve BER performance. Specifically, the encoder is modified so that it does not write information in those bit positions, but inserts randomly chosen dummy bits. If a free-distance error event occurs during decoding, only these positions in the information frame will be affected, so the actual information bits remain uncorrupted.

If we decompose the asymptotic performance bound into spectral lines as in [2], it is clear that the spectral line corresponding to $d_{free} = 6$ has the greatest impact on the performance. By avoiding the bit positions that can be affected by free-distance error events, the modified encoder effectively removes the contribution to the BER represented by the free-distance spectral line. The observed decoded BER should therefore be reduced by the amount reflected in the asymptote.

Fig. 4 compares the simulated BER curve of Fig. 2 to the BER curve obtained when the bit positions corresponding to free-distance codewords are ignored. The simulated BER is reduced by a factor of about 2 relative to the original BER. Also shown is the distance-8 asymptote reflecting the weight-8 codewords listed in Table 1. The plot shows

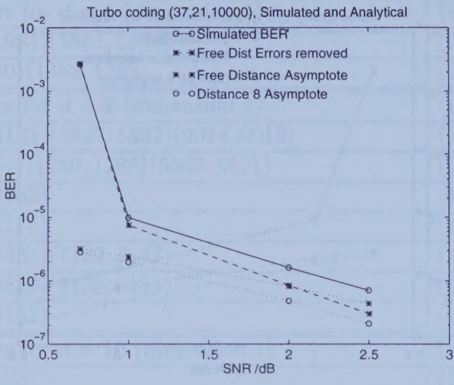


Figure 4: Simulated BER with distance-6 events removed.

the expected changes to the error floor, namely a lowering and change in slope.

The modified encoder incurs a rate loss through this introduction of dummy bit positions. However, the impact is slight for large interleaver size. For the length $N = 10000$ interleaver studied in this paper, the loss is 8 information bits out of 10000, implying a reduction in rate of approximately .04 percent. If applied to the $N = 65536$ turbo code in [2], the modified encoding procedure would discard only 6 information bits out of 65536, causing a decrease in the code rate amounting to less than one hundredth of one percent.

Given a characterization of bit positions affected by other low-distance error events, the encoder can be further modified to avoid writing information into those locations, thereby providing additional improvement to the BER performance. By referring to Table 1, we modified the encoder to avoid positions corresponding to error events of weight 8, in addition to those of weight 6. The performance improvement resulting from this further modification is shown in Fig. 5. The simulated BER is reduced by an additional factor of about 3, yielding a total reduction factor of approximately 7 relative to the original BER. The figure also shows the distance-10 asymptote reflecting the contributions of the weight-10 codewords in Table 1. The rate loss incurred in achieving this reduced BER was slight, amounting to only 34 information bits per frame, or about 0.17 percent.

Finally, the encoder was modified to successively avoid all of the positions affected by the error events with distance 16 or less that we found in our search. Fig. 6 shows the progressive reduction in the error floor, culminating in a BER improvement that is about an order of magnitude at an SNR of 2.5 dB. Note that there is no significant improvement after the removal of the bits corresponding to the distance-10 events. This is due to the existence of distance-12 events not uncovered in the search. The rate penalty incurred by the removal of all of the events of distance 10 or less amounts to only 72 information bits out of 10000, or a reduction of about 0.36 percent. The removal of all of the events of distance 16 or less uncovered in the search amounts to eliminating only 138 information bits out of 10000, for a rate reduction of about 0.7 percent.

We have investigated two variations of the performance improvement scheme proposed above. In the first, we use dummy bits that are known to the decoder and therefore do not need to be transmitted. For example, the dummy bits may be chosen to be all 0's, and the receiver then associates to their bit positions a strong indication that the

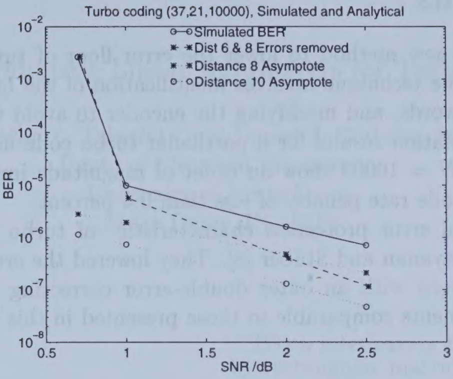


Figure 5: Simulated BER with distance-6 and distance-8 events removed.

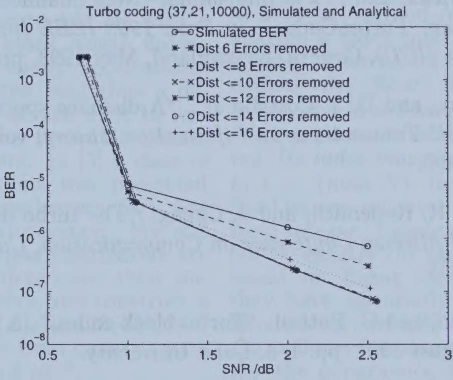


Figure 6: Simulated BER with information weight 2 events of distance 16 or less removed.

transmitted bits were all 0's. The simulation results for this enhanced decoder indicate some improvement in performance, but the gain, only about 0.05 dB, is slight. This approach can be further modified to require that only one information bit position from each information-weight 2 error event be treated as a known dummy bit position, thereby reducing the rate loss.

In a second variation, we attempted to increase the minimum distance of the turbo code by designing an interleaver that avoids the permutations described in (5)-(7). In [6], we applied this criterion in the design of an "S-random" permutation, as introduced by Divsalar and Pollara [7]. The BER results obtained with this constrained "S-random" interleaver were comparable to those shown in Fig. 6. The implementation of this design strategy is straightforward for long interleavers, but for short interleavers, one might expect the constraint on the permutation to limit the "randomness" of the interleaver and, therefore, negatively affect performance gains. It is possible that a combination of the constrained interleaver approach and the dummy bit insertion technique presented in this paper may further improve turbo code performance.

5 Conclusions

We have presented a new method to lower the error floor of turbo codes with sparse distance spectrum. The technique involves identification of the bit positions associated with low-weight codewords, and modifying the encoder to avoid writing information in those locations. Simulation results for a particular turbo code using a pseudo-random interleaver of length $N = 10000$ show an order of magnitude improvement in BER at SNR=2.5 dB with a code rate penalty of less than 0.4 percent.

Note: The unequal error protection characteristic of turbo codes was recognized independently by Narayanan and Stüber [8]. They lowered the error floor by protecting error prone bit positions with an outer double-error correcting BCH code, achieving performance improvements comparable to those presented in this paper.

References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo Codes," in *Proc. 1993 IEEE International Conference on Communication (ICC)*, Geneva, Switzerland, May 1993, pp. 1064–1070.

[2] L. Perez, J. Seghers, and D. J. Costello Jr., "A distance spectrum interpretation of turbo codes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1698–1709, November 1996.

[3] R. J. McEliece, E. R. Rodemich, and J. Cheng, "The turbo decision algorithm," in *Proc. 33 rd Annual Allerton Conference on Communication, Control, and Computing*, October 1995, p. 366.

[4] C. Berrou, S. Evano, and G. Battail, "Turbo-block-codes," in *Turbo coding seminar*, Lund, Sweden, August 1996, pp. 1–8, Lund University.

[5] S. Benedetto and G. Montorsi, "Unveiling turbo codes: Some results on parallel concatenated coding schemes," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 409–428, March 1996.

[6] M. Öberg, A. Vityaev, and P. H. Siegel, "The effect of puncturing in turbo encoders," in *Proceedings of the International Symposium on Turbo Codes & Related Topics*, Brest, France, September 1997, pp. 184–187, ENST de Bretagne.

[7] D. Divsalar and F. Pollara, "Turbo codes for PCS applications," in *Proc. 1995 IEEE International Conference on Communication (ICC)*, Seattle, WA, June 1995, pp. 54–59.

[8] K. R. Narayanan and G. L. Stüber, "Selective serial concatenation of turbo codes'," *IEEE Communications Letters*, vol. 1, no. 5, pp. 136–139, September 1997.

On Deterministic Linear Interleavers for Turbo-Codes*

Oscar Y. Takeshita and Daniel J. Costello, Jr.
Dept. of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556

Summary

The original idea of a turbo-code $C_T(\mathcal{I}_{65536}, (37,21), 1/2)$ used a systematic recursive convolutional code (SRCC) $(37,21)$[1] punctured to rate 1/2 [1] with a random interleaver of block length 65536, prohibitively large for many applications. In [4] a block interleaver, and thus a deterministic interleaver, of size 21 × 19 was shown to produce more favorable results than random interleavers. In [5] a class of non random permutations was presented (we call them linear interleavers $\mathcal{L}_N$), but found to give poor performance. We will show that block and linear interleavers are in a sense equivalent, determine their important design parameters, and construct a simple turbo-code $C_T(\mathcal{L}_{16384}, (23,35), 1/2)$ that is only 1.3dB away from capacity at a Bit Error Rate (BER) of 10^{-5}.

Let $\overline{x} = (x_0, x_1, \ldots, x_{N-1}) \in \{0,1\}^N$. An interleaver $\mathcal{I}_N$ maps $\overline{x}$ to a sequence $\overline{y}$ such that $\overline{y}$ is a permutation of the elements of $\overline{x}$. Let $A = \{0, \ldots, N-1\}$; $\mathcal{I}_N$ can then be defined by the *index mapping function* $d_{\mathcal{I}_N} : A \to A$, $d_{\mathcal{I}_N} : i \mapsto j$, $i, j \in A$, and it can be expressed as an ordered set called the *permutation vector* $\overline{\mathcal{I}}_N = [d_{\mathcal{I}_N}(0), d_{\mathcal{I}_N}(1), \ldots, d_{\mathcal{I}_N}(N-1)]$.

It is known that weight-2 input sequences of the form $\overline{x}_\tau(D) = (1+D^\tau)D^s$, where τ is the cycle-length of the feedback polynomial of the SRCC, are the most problematic for a turbo-code, and that block interleavers can effectively break up bad weight-2 sequences, but not bad weight-4 sequences. A pictorial explanation is given in [5] and [3]. We will give a new algebraic interpretation for this effect.

Block interleavers $\mathcal{B}_N$ are usually defined by rectangular matrices of size $N = m \times n$ that are written row-wise and read column-wise. The *index mapping function* $d_{\mathcal{B}_N}$ of a block interleaver $\mathcal{B}_N$ is of the form $d_{\mathcal{B}_N}(i) \equiv ni + \lfloor i/m \rfloor \pmod N$, $0 \le i < N$. To analyze a block interleaver, we shall use a slightly different interleaver $\mathcal{L}_N$ that "linearizes" the "floor" function $\lfloor\cdot\rfloor$ of $d_{\mathcal{B}_N}(i)$. In [5], the interleaver $\mathcal{L}_N$ is called a *non-random permutation based on circular shifting*. Its *index mapping function* is $d_{\mathcal{L}_N}(i) \equiv ki + v \pmod N$, $0 \le i < N$, where k is a fixed integer relatively prime to N and v is a fixed integer. Interleavers $\mathcal{B}_N$ and $\mathcal{L}_N$ cannot be exactly the same, but they are both based on linear congruences, and hence they have essentially the same properties. Henceforth, we will refer to both $\mathcal{B}_N$ and $\mathcal{L}_N$ as linear interleavers. For an interleaver $\mathcal{L}_N$ the congruence $tk \equiv \tau \pmod N$ is always solvable in t with a unique solution. Hence, an input sequence of the form $\overline{x}_i(D) = (1+D^t)D^q$ to $\mathcal{L}_N$ produces the output $(1+D^\tau)D^s = \overline{x}_\tau(D)D^s$, i.e., it produces a low weight parity sequence for the second component code. By appropriate linear combinations, we can verify that a weight-4 input sequence of the form $\overline{x}_{\text{bad4}} = \overline{x}_i(D) + \overline{x}_i(D)D^\tau$ produces low weight parity sequences in both component codes (the total codeword weight is 12). Moreover, each of the N possible cyclic shifts of $\overline{x}_{\text{bad4}}$ produces the same result.[2] In [2], the following approximate expression for the BER $P_b(e)$ is derived

$$P_b(e) \approx \frac{1}{2}\sum_{w_c} D_{w_c}\,\mathrm{erfc}\left(\sqrt{w_c\frac{R_cE_b}{N_o}}\right), \quad (1)$$

where $D_{w_c} \triangleq \sum_{w_i+w_p=w_c} \frac{w_i}{N} A_{w_i,w_p}$, w_i is the weight of the input sequence, w_p is the weight of the parity sequence, and A_{w_i,w_p} is the number of codewords with input and parity weights w_i and w_p, respectively. If we take only the terms in (1) corresponding to input sequences of the form $\overline{x}_{\text{bad4}}$ we have

---

*This work was supported by NSF Grant NCR95-22939 and NASA Grant NAG5-557.

[1] 37 and 21 are the octal representation of the feedback and feedforward polynomials, respectively.

[2] We suppose that the cyclic shift of the sequence has period N and we disregard edge effects.

$$P_b(e|\overline{x}_{\text{bad4}}) \approx 2\text{erfc}\left(\sqrt{6E_b/No}\right), \qquad (2)$$

which is independent of the interleaver length N and the angular coefficient k, i.e., there is no interleaver gain for this term in the approximation of the BER curve given by (1) and we cannot expect better performance than this asymptote.

We now confirm by simulations previously known results [4] that the performance of linear interleavers is dependent on their angular coefficient; moreover, we establish a new result that for a wide range of interleaver lengths, if we carefully choose the angular coefficient, then we obtain BER curves that have the same asymptote as (2) for sufficiently long interleavers. (See Fig. 1.)[3]

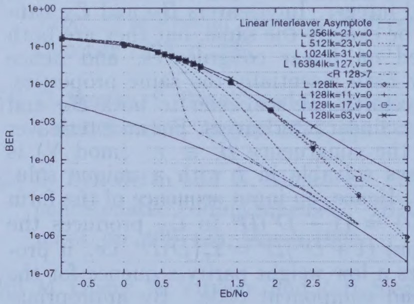


Figure 1: The BER performance of the best linear interleavers found for the turbo-code $C_T(\mathcal{L}_N, (37, 21), 1/2)$. For $N = 128$, we show several k's and an average of random interleavers.

A linear interleaver $\mathcal{I}_N$ with *index mapping function* $d_{\mathcal{I}_N} : i \mapsto j$ can be represented by N lattice points in the (i, j)-plane. We found that a good value for an angular coefficient k is on the order of $\sqrt{N}$, which distributes the lattice points over the plane. The best value is easily found by computing the distance spectrum of the code due to weight-2 input sequences.

In Fig. 1 we show the BER curves, obtained by simulations, for linear interleavers $\mathcal{L}_{128}$ along with the average performance of 7 random interleavers.

It has been observed in the literature [3], that the use of SRCC's with a primitive feedback polynomial in the component codes improves the BER performance (using random interleavers) by increasing the free distance. We now show that similar beneficial effects are also obtained for linear interleavers. The linear interleaver asymptote of the $C_T(\mathcal{L}_N, (23, 35), 1/2)$ turbo-code is estimated by $P_b(e|\overline{x}_{\text{bad4}} \text{ or } \overline{x}_{\text{bad9}}) \approx 2\text{erfc}(\sqrt{10E_b/No}) + 9/2\text{erfc}(\sqrt{9.5E_b/No})$. In Fig. 2 we show simulation results of "primitive" $C_T(\mathcal{L}_N, (23, 35), 1/2)$ turbo-codes. Note the powerful non-random turbo-code with block length 16384 is only 1.3dB away from capacity at a BER of 10^{-5}.

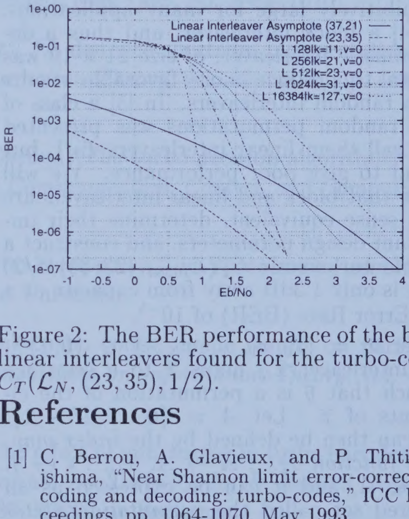


Figure 2: The BER performance of the best linear interleavers found for the turbo-code $C_T(\mathcal{L}_N, (23, 35), 1/2)$.

References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes," ICC Proceedings, pp. 1064-1070, May 1993.

[2] S. Benedetto and G. Montorsi, "Unveiling turbo codes: some results on parallel concatenated coding schemes," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, March 1996.

[3] L. C. Perez, J. Seghers, and D. J. Costello, Jr., "A distance spectrum interpretation of turbo codes," *IEEE Trans. Inform. Theory*, vol. 42, no. 6, November 1996.

[4] A. S. Barbulescu and S. S. Pietrobon, "Interleaver design for turbo codes," *Electronics Letters*, vol. 30, no. 25, pp. 2107-2108, December 1994.

[5] S. Dolinar and D. Divsalar, "Weight distributions of turbo codes using random and nonrandom permutations," JPL TDA Progress Report 42-122, pp. 56-65, August 15, 1995.

---

[3] We use iterative MAP decoding with 18 iterations in all simulations.

An Adaptive Transmission Method for Frequency-Hop Radio Networks

Michael B. Pursley and Clint S. Wilkins
Department of Electrical and Computer Engineering, Clemson University
337 Fluor Daniel Building
Clemson, SC 29634

Abstract

A communication link in a mobile wireless communication network exhibits time variations in both the interference and the propagation loss. Adaptation of the transmitter power and the code rate can greatly improve the throughput and energy efficiency of such a network. The use of adaptive transmission is examined for slow-frequency-hop radio networks in which the radios can adjust the power in the transmitted signal and the rate of the Reed-Solomon code to respond to variations in partial-band interference and propagation loss. A method is proposed for adapting the power and code rate, and the performance of this method is evaluated for a network that employs errors-and-erasure decoding with erasure decisions that are determined by side information from the frequency-hop receiver. The adaptation is based on this side information and on information derived from the decoder.

I. Introduction

A common design approach for a spread-spectrum communication system is to select the transmission parameters according to the worst-case channel conditions. In a frequency-hop communication link with partial-band interference, for example, this approach corresponds to designing the system for the maximum propagation loss and the maximum bandwidth for the partial-band interference that will be encountered in the operation of the system. While such an approach does provide reliable communications for all channel conditions, it is very inefficient in terms of transmitter power and information rate when channel conditions are more favorable. The power is higher than needed when the propagation loss is low, and the code rate is lower than needed when the partial-band interference occupies a very small fraction of the frequency band. If the symbol rate is held constant, a lower code rate results in a longer transmission time for a given message, which in turn results in reduced throughput. If the power is also held constant (or increased), the longer transmission time also results in increased energy consumption. Because the interference and propagation loss vary with time in a mobile wireless network, it is important to adapt the transmitter power and the code rate to match the channel conditions. This leads to higher throughput and greater energy efficiency.

In this paper, Reed-Solomon (RS) codes are employed for error-control in a slow-frequency-hop (SFH) communication network. Side information obtained from the parity-check method [1] is employed to erase individual symbols and unreliable dwell intervals at the output of the SFH demodulator. This side information is used also as one input to the adaptive transmission system which adapts both the transmitter power and the code rate.

We show that there are some scenarios in which it is more effective to change the transmitter power rather than the code rate, and we provide other scenarios in which it is

---

This research was supported by the Army Research Office under grants DAAH04-95-1-0247 and DAAH04-94-G-0154.

more effective to change the code rate. Additional examples are presented in [3]. In some situations it may be necessary to change both the transmitter power and code rate in order to match the transmission to the channel. We find that combined adaptation of power and code rate is necessary to provide reliability and energy efficiency for transmissions in a mobile SFH network.

An algorithm is proposed for adapting the power and code rate to match the interference and propagation loss of the channel. The performance of this method is measured by a normalized throughput, which is essentially the average number of correct symbols at the decoder output per transmitted symbol and per unit energy. With this measure, transmissions are penalized if they use excessive energy or excessive redundancy in order to achieve reliable reception. The performance of the proposed adaptation algorithm is evaluated for a network which employs errors-and-erasure decoding with erasure decisions that are determined by side information from the frequency-hop receiver. The adaptation algorithm relies on this side information and on information derived from the decoder.

II. System and Channel Models

The SFH system employs binary transmission, and the nonbinary symbols for the RS code are formed from sequences of binary symbols. L code words are interleaved to form a packet that is transmitted in n consecutive dwell intervals [2]. The hopping patterns are modeled as sequences of independent, identically distributed random variables, each of which is uniformly distributed on the set of allowable frequencies.

Although the partial-band interference channel is used to illustrate the benefits of adapting the power and code rate in SFH transmissions, similar benefits are expected for channels with SFH multiple-access interference. Partial-band interference is modeled as band-limited white Gaussian noise that occupies a fraction ρ of the band. Its two-sided power spectral density is $\rho^{-1}N_I/2$. Thermal noise with power spectral density $N_0/2$ is also present, so the total power spectral density is $\rho^{-1}N_I/2 + N_0/2$ in a fraction ρ of the band and $N_0/2$ in the remainder of the band.

The parity-check method [1] is used to provide side information in the manner described in [3]. Each code symbol is a parity-check word of length $1+\log_2 n$, which detects any odd number of bit errors. If errors are detected in a received code symbol, the code symbol is erased. The entire dwell interval is erased if errors are detected in more than γ code symbols.

Each of the available codes is an (n, k_j) extended RS code. As in [3], the binary symbols that are transmitted on the channel are orthogonal, equal-energy signals. The transmitted symbol rate is fixed, and the receiver performs optimum bit-by-bit noncoherent demodulation.

The code that is actually used at a given time is selected from a set of two or more codes, and the power level is chosen from a set of two or more levels. The jth code is an (n, k_j) RS code, and the probability of decoding a received word correctly when this code is used is denoted by $P_C(i, j)$ for a transmission which employs the ith power level. Our interest is in a system that adapts both the code rate and the transmitter power. We refer to such a system as an *adaptive transmission system.*

III. Performance Analysis

If an (n, k_j) RS code and the ith power level are employed, the received energy per binary symbol is $E_{bs}(i)$. We assume that N_0, N_I, ρ, and the propagation loss are fixed during the transmission of a packet. The values of N_0, N_I, ρ, and propagation loss during a packet transmission is referred to as the channel state. The *normalized throughput* S, defined as the average number of correct information bits at the decoder output per transmitted binary symbol, is given by

$$S = \frac{k_j}{n + (\log_2 n)^{-1} n} P_C(i,j). \tag{1}$$

The *throughput efficiency* [3] is defined by

$$S_e = \frac{S}{\beta(i)}, \tag{2}$$

where $\beta(i) = E_{bs}(i)/(N_0 + N_I)$. For fixed values of N_0 and N_I, the throughput efficiency represents the normalized throughput per unit energy.

The probability of decoding a received word correctly is

$$\begin{aligned} P_C(i,j) &= \sum_{\ell=0}^{n-k_j} \sum_{m=0}^{m(\ell)} \frac{n!}{(\ell!)(m!)(n-\ell-m)!} (p_t(i))^m (p_e(i))^\ell \\ &\quad \cdot (1 - p_t(i) - p_e(i))^{n-\ell-m}, \end{aligned}$$

where $m(\ell) = \lfloor (n - k_j - \ell)/2) \rfloor$, $p_t(i) = \rho p_t^I(i) + (1-\rho) p_t^N(i)$, and $p_e(i) = \rho p_e^I(i) + (1 - \rho) p_e^N(i)$. Expressions for $p_t^I(i)$, $p_t^N(i)$, $p_e^I(i)$, and $p_e^N(i)$ are given in [3].

IV. An Adaptive Transmission Algorithm and Its Performance

The algorithm used for adapting the code rate and power level is as follows. The receiver determines the value of the EE metric for each dwell interval. The EE metric for a group of code symbols is the number of symbol erasures plus twice the number of symbol errors. The EE metric was introduced in [4] where it is employed to provide link-quality information to be used in routing decisions in a mobile packet radio network. For each packet, the number Y of dwell intervals for which the value of the EE metric exceeds λ is determined. If $Y > \tau$ the highest power level and the lowest code rate are selected for use in the next packet transmission. If $Y \leq \tau$ the receiver computes $\hat{p}_{ave}$, which is defined as

$$\hat{p}_{ave} = \frac{1}{n - Y} \sum_{i=1}^{n-Y} \hat{p}_i,$$

where $\hat{p}_1, \hat{p}_2, \ldots,$ and $\hat{p}_{n-Y}$ are the values of the EE metric for the $n - Y$ dwell intervals in which the EE metric does not exceed λ. A code rate is selected based on the value of Y and a power level is selected based on the value of $\hat{p}_{ave}$.

To illustrate how the code rate and power level are selected, consider an adaptive transmission system that selects a code from among a set of two codes and a power level from among a set of two power levels. If $Y \leq \tau$, then Y is compared to a single threshold τ_c, where $\tau_c < \tau$. The lowest code rate is selected if $Y > \tau_c$, and the highest code rate is selected if $Y \leq \tau_c$. The power level is determined in the following way: if $\hat{p}_{ave} \leq \tau_1$ the power is decreased if possible; if $\tau_1 < \hat{p}_{ave} \leq \tau_2$ the power level is unchanged; otherwise the power is increased if possible. By using a greater number of thresholds, this idea can be extended easily to radios that use more than two code rates or more than two power levels .

If a received word is not decoded correctly, it is not possible to determine which symbols are in error in the received word, and the receiver must estimate the number of errors. It does know the number of erasures, however. In most applications, the probability of not decoding correctly is small, and simple estimates should suffice. For the purposes of analysis, the adaptations of the code rate and power level are based on the actual value of the EE metric rather than on an estimate.

Response	Throughput Efficiency	Probability of Response
No change	0.007	0.024
Increase power by 3 dB	0.012	0.000
Increase power by 3 dB and change code to (32,12)	0.097	0.009
Change code to (32,12)	0.186	0.967

Table 1. Conditions: E_{bs}/N_0=9 dB, E_{bs}/N_I=3 dB, $\rho = 0.1$. Transmission using $P_{nom}/2$, (32,24) RS code.

All of the results are for systems with the number of code words per packet $L = 10$ and the threshold $\gamma = 3$ for the parity-check method. For each set of channel conditions there is a value of γ that maximizes the probability of decoding correctly. Fortunately, γ=3 provides performance that is nearly optimum over a wide range of channel conditions. The available codes for the examples we present are from the set $\{(32,24),(32,12)\}$, and the two power levels are a nominal power P_{nom} and one-half the nominal power $P_{nom}/2$. In all of the examples, $\lambda = 5, \tau = 19, \tau_c = 7, \tau_1 = 0.15$, and $\tau_2 = 1.40$. These values give good performance over a wide range of channel conditions as demonstrated in the examples that follow, but we make no claims regarding their optimality.

Suppose the transmitter power level is $P_{nom}/2$ and the corresponding values of E_{bs}/N_0 and E_{bs}/N_I are 9 dB and 3 dB, respectively. Suppose also that $\rho = 0.1$ and a (32,24) code is used. The resulting throughput efficiency is 0.359. If ρ changes to 0.4, the throughput efficiency is reduced to 0.007, so the adaptive transmission system should respond by adapting the code rate and power level to provide the largest possible throughput efficiency for the two available codes and the two allowed power levels. In Table 1 all possible system responses are shown along with their corresponding throughput efficiencies. The probability of each system response is also presented. The combination of power level and code rate that provides the best throughput efficiency for a given channel state is referred to as the desired state of the adaptive transmission system. Any other state of the adaptive transmission system is referred to as an undesired state. It is clear that use of $P_{nom}/2$ and a (32,12) code is the desired state of the adaptive transmission system in this situation. The adaptive transmission algorithm selects the desired state with probability 0.967.

Now suppose the transmitter uses power level P_{nom}, and the corresponding values of E_{bs}/N_0 and E_{bs}/N_I are 9 dB and 3 dB, respectively. If $\rho = 0.4$ and a (32,12) code is used, the system's throughput efficiency is 0.186. If E_{bs}/N_0 changes to 12 dB and E_{bs}/N_I changes to 6 dB, the throughput efficiency drops to 0.097. The data in Table 2 show the throughput efficiency that results from all possible system responses. The probability of each system response is also presented. Note that the system chooses the desired state with probability 0.942.

The data in Table 1 illustrate a situation in which it is most effective to change the code rate rather the power level while the data in Table 2 show a situation in which it is most effective to change to power level rather than the code rate. These examples corroborate the conclusion in [3] that it is necessary to adapt both the transmitter power and the code rate. The adaptive transmission algorithm selects the desired state with high probability in each of these examples.

The data in Table 3 show that one of the undesired states is selected with probability

Response	Throughput Efficiency	Probability of Response
No change	0.097	0.032
Change code to (32,24)	0.012	0.001
Decrease power by 3 dB and change code to (32,24)	0.007	0.025
Decrease power by 3 dB	0.186	0.942

Table 2. Conditions: E_{bs}/N_0=12 dB, E_{bs}/N_I=6 dB, $\rho = 0.4$. Transmission using P_{nom}, (32,12) RS code.

Response	Throughput Efficiency	Probability of Response
No change	0.103	0.036
Change code to (32,24)	0.057	0.010
Decrease power by 3 dB and change code to (32,24)	0.020	0.202
Decrease power by 3 dB	0.202	0.752

Table 3. Conditions: E_{bs}/N_0=11 dB, E_{bs}/N_I=6 dB, $\rho = 0.3$. Transmission using P_{nom}, (32,12) RS code.

Response	Throughput Efficiency	Probability of Response
No change	0.020	0.141
Increase power by 3 dB	0.057	0.030
Increase power by 3 dB and change code to (32,12)	0.103	0.118
Change code to (32,12)	0.202	0.711

Table 4. Conditions: E_{bs}/N_0=8 dB, E_{bs}/N_I=3 dB, $\rho = 0.3$. Transmission using $P_{nom}/2$, (32,24) RS code.

0.202. Suppose the adaptive transmission system selects a (32,24) code and the power level $P_{nom}/2$. Data are presented in Table 4 for this new state of the system. In this case the desired state is selected with probability 0.711. In Figure 1 a state diagram with all possible states of the adaptive transmission system is presented. The only transitions shown are those to the desired state, and these transitions are labeled with their associated probabilities. The state diagram shows that the desired state is by far the most probable state of the adaptive transmission system regardless of its previous state.

V. Conclusions

In this paper, we propose an algorithm for adaptating of the power level and code rate in a slow-frequency-hop communication system. Results which demonstrate the potential of the algorithm are presented for a simple adaptive transmission system with two code rates and two power levels.

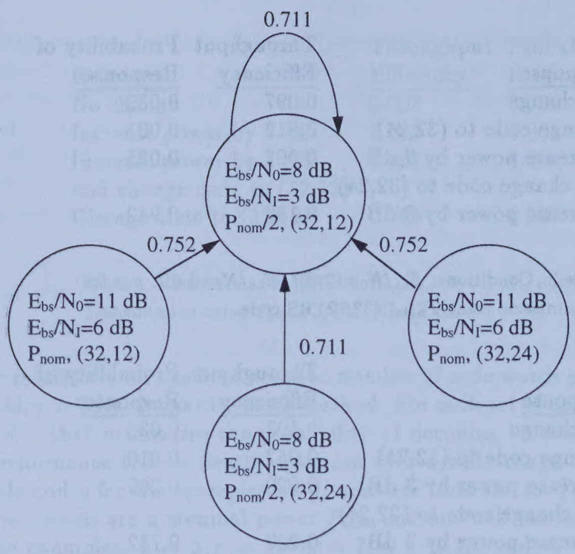

0.711
E_{bs}/N_0=8 dB
E_{bs}/N_I=3 dB
P_{nom}/2, (32,12)
0.752
0.752
E_{bs}/N_0=11 dB
E_{bs}/N_I=6 dB
P_{nom}, (32,12)
0.711
E_{bs}/N_0=11 dB
E_{bs}/N_I=6 dB
P_{nom}, (32,24)
E_{bs}/N_0=8 dB
E_{bs}/N_I=3 dB
P_{nom}/2, (32,24)

Fig. 1. State diagram of adaptive transmission system for $\rho = 0.3$ and fixed values of N_0, N_I, and path loss.

References

[1] A.W. Lam, and D.V. Sarwate, "A comparison of two methods for generation of side information in frequency-hop spread-spectrum multiaccess communications," *Conf. on Inform. Sciences and Systems*, pp. 426-431, March 1987.

[2] M.B. Pursley, "Reed-Solomon codes in frequency-hop communications," in *Reed Solomon Codes and Their Applications*. Piscataway, NJ: IEEE Press, 1994, pp. 150-174.

[3] M.B. Pursley and C.S. Wilkins, "Adaptive transmission for Frequency-Hop Communications with Reed-Solomon Coding," *IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, pp. 866-869, Aug. 1997.

[4] M. B. Pursley and H. B. Russell, "Network protocols for frequency-hop packet radios with decoder side information," *IEEE Journ. Sel. Areas Commun.*, vol. 12, pp. 612-621, May 1994.

Synchronous Coherent OFDM for Broadband Cellular Radio Systems

Kevin L. Baum
Chicago Corporate Research Laboratories
Motorola
1301 E. Algonquin Road, Room 2928
Schaumburg, IL 60196 USA

Abstract

This paper investigates a new air interface concept for future OFDM-based cellular systems called Synchronous Coherent Orthogonal Frequency Division Multiplexing (SC-OFDM). SC-OFDM is based on a combination of synchronized cells, orthogonal pilot codes, and a cyclic extension which is large enough to absorb both delay spread and intercell propagation delay. The primary advantages of SC-OFDM are that it mitigates the effect of co-channel interference on desired signal channel estimates and facilitates the direct measurement of the channel response of individual co-channel interfering signals. These measurements provide channel state information and enable the use of interference-mitigating detection techniques such as MMSE diversity combining or joint detection. The performance of an SC-OFDM downlink air interface is evaluated with computer simulations.

1.0 Introduction

In order to serve the growing demands of wireless communication users and provide new services such as wireless video conferencing and Internet access, high data rate cellular communication systems are currently being studied around the world. In Europe, for example, efforts are underway to develop a standard for a next-generation wireless system called the Universal Mobile Telecommunications System (UMTS). A goal of this system is to support a wide range of data rates including low-rate services (e.g., voice, paging) and high-rate services of up to 2 Mbps. Several physical layer technologies are being studied including TDMA, CDMA, and Orthogonal Frequency Division Multiplexing (OFDM), along with various combinations of these technologies.

OFDM has been studied or adopted for several applications [1,2] in the past, including Digital Audio Broadcast (DAB) [1] and Asymmetric Digital Subscriber Line (ADSL). One of the most recognized benefits of OFDM is that complex channel equalization can be avoided even in very high data rate systems through the use of a cyclic extension (or guard interval) between symbols. Moreover, modulation and demodulation can be efficiently implemented using FFT processing [3].

This paper investigates an air interface concept for OFDM-based cellular systems called Synchronous Coherent Orthogonal Frequency Division Multiplexing (SC-OFDM). The SC-OFDM air interface concept is being studied as a possible enhancement for OFDM-based pico-, micro-, and small-cell macrocellular systems. SC-OFDM mitigates the effect of co-channel interference on desired signal channel estimates and facilitates the direct measurement of the channel response of individual co-channel interfering signals. These measurements provide channel state information for soft-decision decoding and channel quality evaluations. Moreover, these measurements can be used in conjunction with interference-mitigating detection methods to suppress co-channel interfering signals.

The paper is organized as follows. After the introduction, the second section reviews the basic aspects of OFDM. The third section describes the SC-OFDM concept in general and details the techniques used to structure the co-channel interference appropriately. The fourth section provides an example SC-OFDM downlink air interface design. Applicable system parameters are given along with reasons for their selection. Simulation results are presented for a two-antenna receiver with minimum mean square error (MMSE) diversity combining. The final section summarizes the characteristics and performance of SC-OFDM.

2.0 OFDM

OFDM transmits M data symbols per baud interval on M distinct subcarrier frequencies. Let the data symbols be denoted as $x(m,n)$, where m is the subcarrier index and n is the baud interval index. The total baud interval duration is specified as $T_s = (t_s + \Delta)$, where $(t_s)^{-1}$ is the subcarrier spacing in Hz, and Δ is the length of the cyclic extension. The n^{th} baud interval occupies a time interval $(nT_s \le t \le (n+1)T_s)$ which includes a cyclic extension of length Δ followed by the length t_s information portion of the OFDM signal. The cyclically extended transmitted signal waveform for the n^{th} baud interval can be written in complex baseband representation as

$$s_n(t) = \sum_{m=0}^{M-1} x(m,n)\, e^{\frac{j2\pi m(t-\Delta-nT_S)}{t_s}} \qquad \text{for} \quad nT_S \le t \le (n+1)T_S. \qquad (1)$$

If this signal is transmitted over an ideal, noiseless channel, the data symbols of the n^{th} baud interval can be recovered as

$$x(m,n) = \frac{1}{t_s} \int_{nT_S+\Delta}^{(n+1)T_S} s_n(t)\, e^{\frac{-j2\pi m(t-\Delta-nT_S)}{t_s}}\, dt \qquad \text{for} \quad 0 \le m \le (M-1). \qquad (2)$$

In practice, the OFDM signal synthesis in the transmitter and analysis in the receiver are efficiently implemented with DFT or FFT processing [3]. If we assume that the cyclic extension is longer than the channel impulse response and that the channel is time-invariant over a baud interval, the corrupted received data at the DFT output for the m^{th} subcarrier and n^{th} baud interval can be modeled as

$$y(m,n) = x(m,n)g(m,n) + z(m,n) \qquad (3)$$

where $z(m,n)$ represents noise and other additive distortion such as co-channel interference, and $g(m,n)$ is the Fourier transform of the channel impulse response evaluated at the frequency of the m^{th} subcarrier during the n^{th} baud interval. In the following, $g(m,n)$ is referred to simply as the channel response. As this equation indicates, a multipath channel does not cause inter-symbol or inter-carrier interference (ISI and ICI) in a cyclically extended OFDM signal as long as Δ is longer than the channel impulse response. The effect of the multipath channel is reduced to a set of multiplicative coefficients which can be measured in the receiver if known (pilot) symbols are periodically included in the transmitted symbol stream.

3.0 The SC-OFDM Concept

SC-OFDM is an extension of OFDM which involves the incorporation of the following concepts into an OFDM-based cellular system:

1. *Synchronized Cells* - Signal transmissions from co-channel cells are time-synchronized. This is done to align the baud intervals of the co-channel transmitters. Base stations must lock to a common clock to achieve precise downlink timing synchronization. For uplink transmissions, a subscriber unit can lock to the base station timing reference using a closed-loop timing advance procedure.

2. *Absorption of Propagation Delay with a Cyclic Extension* - The cyclic extension, which is normally made only as long as the expected channel impulse response duration, is lengthened in SC-OFDM to accommodate both the channel impulse response duration and intercell propagation delay. Since signal transmissions in co-channel cells are time-synchronized, co-channel signals emanating from different transmitters will arrive at a receiver at times which differ only by the intercell propagation delay. In cells of moderate size, this intercell propagation delay will typically be much smaller than T_s, and can be absorbed by increasing Δ. If Δ is sufficiently large to cover the combined channel impulse response duration and propagation delay, the received signal at the m^{th} subcarrier and the n^{th} baud interval can be modeled as the weighted sum of the desired and co-channel interfering data symbols as follows:

$$y(m,n) = x_0(m,n)g_0(m,n) + \sum_{j=1}^{N_j} x_j(m,n)g_j(m,n) + v(m,n) \qquad (4)$$

where $x_0(m,n)$ and $g_0(m,n)$ are associated with the desired signal, $x_j(m,n)$ and $g_j(m,n)$ are associated with the j^{th} co-channel interfering signal for $j \geq 1$, N_j is the total number of interferers, and $v(m,n)$ is additive white Gaussian noise (AWGN).

3. *Intercell Reference Signal Orthogonality* - To facilitate coherent detection and other more advanced interference mitigating detection methods, it is useful to be able to measure the channel response of the desired signal and each interferer. Moreover, when measuring the channel response of the j^{th} signal, the contribution of the other signals should be suppressed to provide a more accurate channel response measurement. SC-OFDM supports the measurement of the channel responses of the desired and interfering signals with a set of orthogonal reference signals. A set of orthogonal pilot codes is used to provide the orthogonal reference signals. A pilot code is a sequence of pilot symbols selected from a set of orthogonal sequences. A sequence length of L provides L orthogonal pilot codes which can be allocated to L different co-channel cells. A pilot code is denoted as a vector $\underline{w}_j$ where the elements, or symbols of the pilot code are denoted $w_{i,j}$. Thus, the j^{th} pilot code of the L orthogonal pilot codes is written as $\underline{w}_j = [w_{0,j}\ w_{1,j} \ldots w_{L-1,j}]^T$ where the superscript T denotes the vector transpose. The orthogonality of the pilot codes is assured by selecting the codes such that $\underline{w}_j^H \underline{w}_k = 0$ for $j \neq k$, where the superscript H denotes the Hermitian transpose. If each transmitter places the L symbols comprising a pilot code in the same adjacent baud intervals on the m_p^{th} subcarrier, we have $x_j(m_p, n_p + i) = w_{i,j}$ for $0 \leq i, j \leq (L-1)$ where n_p is the baud

interval index of the first pilot code symbol. Let $|w_{i,j}| = 1$ for convenience. Then using the model of eq. (3), if each g_j is constant over the pilot code duration and the AWGN is negligible, it can be shown that

$$g_j(m_p, n_p) = \frac{1}{L}\sum_{i=0}^{L-1} y(m_p, n_p + i) w_{i,j}^{*} \qquad \text{for } 0 \le j \le (L-1). \tag{5}$$

This equation illustrates how the channel response of each co-channel signal can be measured over a pilot code without interference from the remaining co-channel signals.

4. *Common Transmission Format* - A common transmission format is used among the co-channel cells to insure that the pilot codes of co-channel transmitters are located in the same group of time-frequency positions. Furthermore, the common transmission format should include slotted or bursted transmission. A slot consists of one or more baud intervals, one or more OFDM subcarriers, and one or more instances of a pilot code. To avoid discontinuities between the interference on pilot codes and data symbols, slots should have the same dimensions in each cell and all interferer channel response measurements should only make use of pilot codes contained within the slot.

4.0 SC-OFDM Downlink Configuration

This section evaluates the SC-OFDM concept for a downlink air interface application. A set of four orthogonal pilot codes, each of length four, was selected to facilitate the measurement of the channel response of four separate signals. In this case, the four signals include the desired signal and three co-channel interfering signals. The use of length four pilot codes is motivated by the geometry of the single-cell, three-sector frequency reuse pattern shown in Figure 1. The available bandwidth is divided into three groups with each of the groups allocated to a different sector within each cell. A subscriber unit, denoted as a small oval in Figure 1, is assumed to be located in the shaded sector of the center cell, also known as the desired sector. The shading in the surrounding cells indicates which sectors within those cells use the same frequency group as the desired sector. As indicated by the arrows, this model limits the number of primary interfering signals within the first ring of surrounding cells to three. Although this is an idealized model, it is useful to illustrate the SC-OFDM air interface concept. In general, the length of the pilot code should be greater than the number of strong co-channel interferers. Otherwise, the capability to individually measure the channel response of interferers will be impaired.

Figure 1 also illustrates pilot code allocations for the given cell/sector reuse plan. The desired sector is allocated pilot code 0, denoted $\underline{w}_0$, and the primary interfering sectors are allocated the remaining three pilot codes, $\underline{w}_1$, $\underline{w}_2$, and $\underline{w}_3$. This allocation scheme can be extended such that all cells and sectors, when viewed in conjunction with their three respective primary interferers, utilize a complete set of length four pilot codes.

4.1 OFDM SIGNAL PARAMETERS

The OFDM signal parameters selected for the example downlink are shown in Table 1. A downlink bandwidth allocation of 15 MHz was assumed, with each of the three sectors in each cell allocated a 5 MHz portion.

Table 1: OFDM Signal Parameters for Example Downlink

Parameter	Value
Information Portion of Symbol (t_s)	80 μs
Cyclic Extension Duration (Δ)	20 μs
Total Baud Duration, Baud Rate (T_s, $\frac{1}{T_s}$)	100 μs, 10 kbaud
Subcarrier Spacing ($\frac{1}{t_s}$)	12.5 kHz
Bandwidth per Sector	5 MHz
Number of Subcarriers per Sector	374
IFFT/FFT Size	512
Forward Error Correction	rate-1/2 conv., 64 state
Modulation	QPSK

Considering delay spreads of a few microseconds and a propagation delay of about 3.3 μs per km, the cyclic extension duration of 20 μs is reasonable for cell sizes up to 1 km in radius. A portion of the cyclic extension can also be used for smooth ramp-up and ramp-down allowing for faster spectral sidelobe rolloff. As with other OFDM systems, the cyclic extension is discarded in the receiver prior to demodulation. Discarding this part of the received signal energy causes about 1 dB of degradation in signal-to-noise ratio performance as compared to an ideal single-carrier QPSK system.

4.2 SLOT AND FRAME FORMAT

The slot format for the SC-OFDM downlink is shown in Figure 2. A slot is 11 subcarriers wide (137.5 kHz) and 6 baud intervals long (0.6 ms). Pilot codes, each of length 4, are placed on subcarriers 1, 4, 8, and 11. This slot format assumes that the channel does not change drastically over the 0.6 ms slot duration. Interpolation is used across the subcarriers to track the channel response variations due to frequency selective fading.

Each transmission frame encompasses 374 subcarriers (~4.7 MHz) by 100 baud intervals (10 ms) and can hold up to 34 slots in frequency and 16 slots in time. Sixteen slots in time occupy 9.6 ms. The remaining 0.4 ms (4 baud intervals) are reserved for the transmission of synchronization signals to support receiver synchronization. These 4 baud intervals can be shared by 4 cells based on the length 4 pilot code allocation shown earlier. By coordinating transmission of the synchronization signals according to the pilot code allocation plan, the amount of co-channel interference on a synchronization signal is reduced significantly. Moreover, a receiver is able to monitor the synchronization signals from the primary co-channel interfering cells.

The aggregate information rate for all slots in the frame is

$$\frac{50\,\text{bits/slot}\cdot(34\cdot 16\,\text{slots})}{10\,\text{ms}} \cong 2.7\,\text{Mbps}$$

which can be shared by multiple users or dedicated to a single user. The data rate provided to a particular user is determined by the number of slots assigned to the user. Since all slots originate from the same transmitter, they can be spread across the frequency band to improve the frequency diversity effect. Moreover, a slow frequency hopping (SFH) capability is provided by periodically re-ordering the slots across the 374 subcarrier frequency band after either every slot or every second slot. As with other coded OFDM systems, the use of SFH in combination with coding and interleaving enables the system to realize a powerful frequency diversity effect. Slots which are not allocated to a user are not transmitted in order to reduce the amount of co-channel interference that is generated.

The simulation results to be presented in Section 4.3 indicate the potential to support single-cell frequency reuse when MMSE diversity reception is used. If single-cell frequency reuse can be supported, the 2.7 Mbps information rate translates to an efficiency of

$$\frac{2.7\,\text{Mbps/sector}\cdot 3\,\text{sectors/cell}}{1\,\text{cell}\cdot 15\,\text{MHz/cell}} = 0.54\ \text{bps/Hz/cell}$$

4.3 PERFORMANCE SIMULATIONS

The performance of the SC-OFDM downlink air interface with 2 antenna diversity reception has been simulated under various channel conditions. For the simulation results presented here, ideal frequency and symbol timing synchronization were used, and the decoded bit error rate (BER) was measured over 400,000 information bits for a user with a 65 kbps data rate. Combining the rate-1/2 forward error correction scheme detailed above with a (30,5) convolutional interleaver yielded an interleaving delay of 30 ms. Digital frequency hopping was employed every 1.2 ms to provide frequency and interference diversity effects. Co-channel interfering signals were assumed to be independently faded from the desired signal. A two element antenna array was modeled at the receiver and each element in the array was assumed to have its own independently faded channel. The signals received by each element were combined using a minimum mean square (MMSE) algorithm, which is based on the measured channel responses of the desired and interfering signals. Table 2 summarizes the parameters common to each of the simulation runs.

Table 2: Parameters Common to All Simulation Runs

Item	Value
Frequency Synchronization	Ideal
Symbol Timing Recovery	Ideal
User Information Rate	65 kbps
Forward Error Correction	rate-1/2 conv., 64 state
Interleaver Size, Delay	(30,5) conv., 30 ms
Decoding	Soft-Decision
Modulation	QPSK
Frequency Hopping Dwell Period	1.2 ms
Receive Antenna Elements	2
Antenna Combining Method	MMSE

In general, the use of an *N* element MMSE receive antenna array supports the suppression of *N*-1 co-channel interferers. Therefore, the two-element array can effectively suppress a single dominant co-channel interferer.

Figures 3 and 4 show performance versus the *energy per QPSK symbol-to-noise power spectral density ratio*, E_s/N_0. In all cases, each of the two diversity branches is given the same E_s/N_0. Figure 3 shows performance in static additive white Gaussian noise (AWGN), flat faded, and 1 μs RMS delay spread channels. For the flat faded and delay spread channels, a Doppler rate of 50 Hz was assumed. Performance in the fading channel is seen to improve when moderate delay spread is introduced. This trend indicates the benefit of frequency diversity.

Figure 4 shows performance in a 1 μs RMS delay spread channel for Doppler rates of 20, 50, and 240 Hz. At a carrier frequency of 2 GHz, these Doppler rates correspond to speeds of approximately 11, 27, and 130 kph, respectively. Despite the fact that the channel response measurement and detection algorithms in the receiver assume that the channel is time-invariant over a slot, the curves of Figure 4 indicate that performance does not degrade significantly at high Doppler rates.

Figure 5 shows performance in co-channel interference environments. Decoded BER curves are plotted against total carrier-to-interference (C/I) ratio. For each of these simulations, the RMS delay spread was fixed at 1 μs, the Doppler rate was set to 50 Hz, and the E_s/N_o was fixed at 10 dB. Although it is common to use a much higher value for E_s/N_o when evaluating BER performance versus C/I, the conservative E_s/N_o value of 10 dB accommodates additional interference and/or implementation degradations which are not specifically modeled in the simulation. Such additional interference may arise either from the second ring of co-channel cells or from adjacent channel signals. All co-channel interferers were assumed to have the same slot allocations and frequency hopping pattern as the desired signal. This model does not provide any interference diversity benefit, effectively simulating a worst-case 100% loading condition.

Three BER curves are presented in Figure 5, each detailing performance for a different co-channel interference model. The first model assumes that all co-channel interference comes from a single base station with a pilot code that is orthogonal to the pilot code of the desired base station. The second and third models assume three interfering base stations, which in conjunction with the desired base station, use a mutually orthogonal set of pilot codes. The second model has three unequal power interferers, with a single dominant interferer, a second interferer 4 dB down from the dominant one, and a third interferer 8 dB down from the dominant one. The third model has three equal power interferers. Note that the simulation results for the single co-channel interferer model had no decoded bit errors. As mentioned earlier, a two-element array is able to suppress a single dominant co-channel interferer. The decoded BER increases as the number of interferers increases to three, but good performance is still obtained at low C/I values.

Figure 6 further details the capability of the system to suppress a dominant co-channel interferer. Performance of the SC-OFDM downlink is shown in a Rayleigh faded AWGN environment. The RMS delay spread of the channel was set to 1 μs and the maximum Doppler rate was fixed at 50 Hz. For the noise-only curve in Figure 6, no co-channel interference was introduced. The other curves show performance under these same conditions but with the addition of a single co-channel interferer using an orthogonal pilot code and received with a C/I of 0 dB. For the second curve, the estimate of the interferer's channel response is disregarded in the diversity combining process. To this extent, the co-channel interference has minimal effect on the desired channel estimates (because of the orthogonal pilot code), but directly interferes with the desired data symbols. Treating the co-channel interference in this manner amounts to maximal-ratio combining of the signals received on the two antenna elements. For the third curve, the interferer's channel estimate is used in conjunction with MMSE combining to suppress the effect of co-channel interference on the desired data symbols. This curve, with a much steeper slope than the max-ratio curve, illustrates the interference suppression capability of SC-OFDM with MMSE diversity combining.

5.0 Discussion and Conclusions

Although SC-OFDM has several potential benefits in strong co-channel interference cellular scenarios, some of its important design constraints should be noted. Since SC-OFDM relies on the cyclic extension to absorb both delay spread and propagation delay, the system parameters must be carefully chosen to minimize the cyclic extension overhead. The cell sizes should be as made as small as possible to minimize propagation delay and delay spread. When SC-OFDM is applied in systems with very large cells, pilot code orthogonality is difficult to maintain and performance will likely be closer to that of a conventional coherent OFDM system that employs a high density pilot symbol pattern. SC-OFDM is therefore seen as providing an enhanced air interface mode for deployments with high user densities, particularly when high data rates of up to 2 Mbps are needed. Also, the requirement that the channel response must not vary significantly over a pilot code must be accounted for when selecting the fundamental signal parameters and the SC-OFDM slot structure. If the channel response varies over the pilot code, orthogonality will be disturbed. Finally, for uplink applications, it is difficult to maintain precise time synchronization among co-channel transmitting subscriber units. A precise closed-loop timing advance scheme must be employed to enable the use of SC-OFDM in uplink applications.

The two-antenna MMSE diversity combining scheme used in the SC-OFDM simulations provides good performance and interference suppression, but the added complexity of the second receiver may be undesirable in some situations. In these cases it is possible to obtain performance gains in a single antenna receiver through the use of joint detection. Since there is no ISI on the desired or interfering signals with SC-OFDM, the joint detector is a simple memoryless device. Joint detection of this type is fairly simple and effective if there is only a single dominant interferer. Results for SC-OFDM with this type of joint detection will be shown in a future paper.

Overall, the results of this study indicate that the SC-OFDM air interface concept merits further consideration for future broadband cellular systems which operate in small-cell, interference-limited environments. The interference measurement capability can be used in several ways including soft-decision weighting and channel quality monitoring. Moreover, when the interference measurement capability is combined with advanced detection methods, the impact of co-channel interference can be reduced significantly.

Acknowledgement

The author would like to thank Nikhil Nadgauda for developing the OFDM link simulator and contributing to this work while he was employed by Motorola.

References

[1] J. C. Rault, D. Castelain, and B. Le Floch, "The Coded Orthogonal Frequency Division Multiplexing (COFDM) Technique and its Application to Digital Radio Broadcasting Toward Mobile Receivers," *Globecom '89*, pp. 428-432, Nov. 1989.

[2] J. A. C. Bingham, "Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come," *IEEE Communications Magazine*, vol. 28, pp. 5-14, May 1990.

[3] S.B. Weinstein, P.M. Ebert, "Data Transmission by Frequency Division Multiplexing Using the Fast Fourier Transform," *IEEE Transactions on Communication Technology*, vol. 19, no. 5, pp. 628-634, May 1971.

[4] M. Russell and G. L. Stuber, "Interchannel Interference Analysis of OFDM in a Mobile Environment," *IEEE Vehicular Technology Conference Proceedings*, pp. 820-824, July 1995.

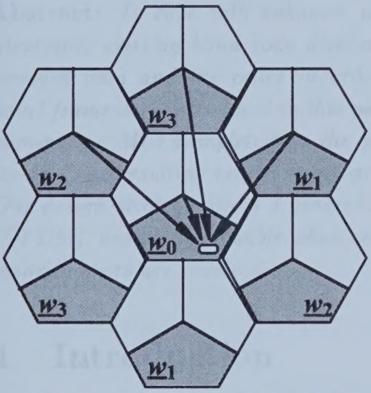


Figure 1: Pilot Code and Co-channel Interference Distribution for Single-Cell Three-Sector Reuse

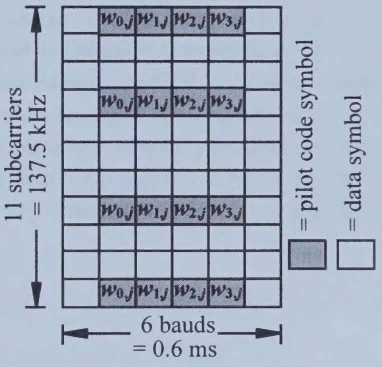

$w_{0,j}$ $w_{1,j}$ $w_{2,j}$ $w_{3,j}$
11 subcarriers = 137.5 kHz
= pilot code symbol
= data symbol
6 bauds = 0.6 ms

Figure 2: SC-OFDM Slot Format

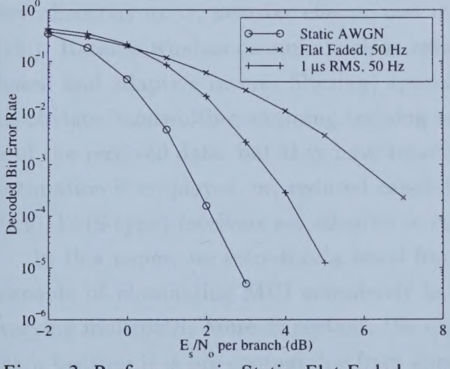


Figure 3: Performance in Static, Flat Faded, and Delay Spread Channels (50 Hz Doppler).

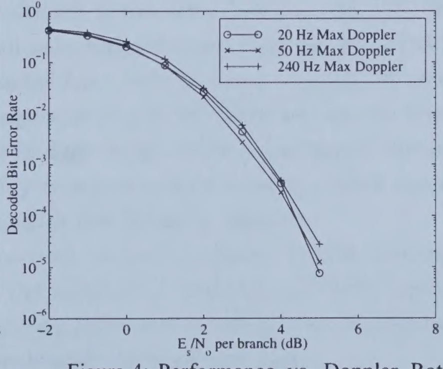


Figure 4: Performance vs. Doppler Rate at 1μs RMS Delay Spread.

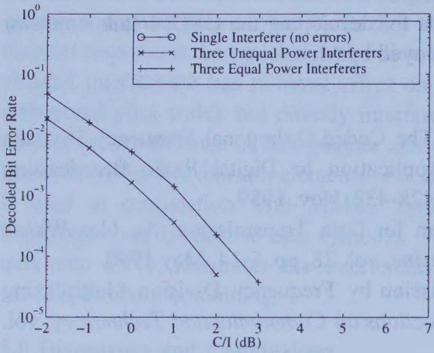


Figure 5: C/I Performance with E_s/N_o = 10 dB, 1μs RMS Delay Spread, 50 Hz Doppler.

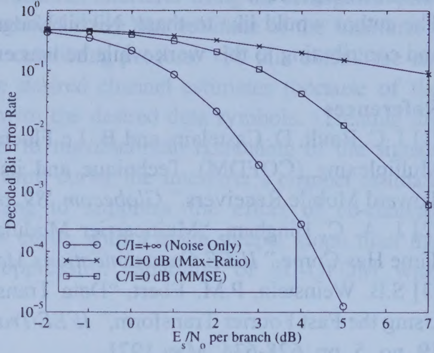


Figure 6: Interference Suppression in SC-OFDM with MMSE Combining.

Designing user codes for MUI elimination and unknown multipath mitigation in asynchronous CDMA systems

Georgios B. Giannakis[1] and *Anna Scaglione*[2]

[1]Dept. of Electrical Engineering, University of Virginia, Charlottesville, VA 22903-2442, USA
e-mail: `georgios@virginia.edu`, Tel: (804) 924-3659, Fax: (804) 924-8818

[2]Infocom Dept., Univ. of Rome "La Sapienza", via Eudossiana 18, 00184 Roma, ITALY
`annas@infocom.ing.uniroma1.it`, Fax: (+39) 6-4873300

Abstract: *To cope with unknown multipath propagation in wireless multiuser communications, existing blind code division multiple access (CDMA) approaches rely on the received data and the codes in order to suppress multi-user interference (MUI). The novel framework introduced in this paper yields low complexity CDMA systems capable of eliminating MUI completely in the presence of unknown and even rapidly varying multipath. The resulting receivers are pre-computed from appropriately designed user codes. Our design strategy offers a generalization of orthogonal frequency division multiplexing (OFDM), especially valuable when carrier frequency errors, Doppler effects, and different priority users are involved.*

1 Introduction

The recognition of CDMA as a promising multiple access scheme was followed by efforts to design simple linear receivers (zero-forcing or mean-square error) capable of handling asynchronous users, near-far effects, and multipath propagation (see e.g., [6], [11], [4], [10]). Because wireless communications entail unknown multipath channels, blind (SVD based and adaptive inverse filtering) approaches have been proposed recently, in order to obviate bandwidth-consuming training sequences [9], [5], [8]. They rely on the codes and the received data, but they have relatively high complexity when subspace channel estimation is employed, or, reduced capability to suppress MUI especially when simple (e.g., LMS-type) receivers are adopted to equalize fast changing channels.

In this paper, we introduce a novel framework for low-complexity CDMA receivers capable of eliminating MUI completely in the presence of *unknown* and even rapidly varying multipath. More important, the resulting receiver does not rely on the received data because it is pre-computable from appropriately designed user codes.

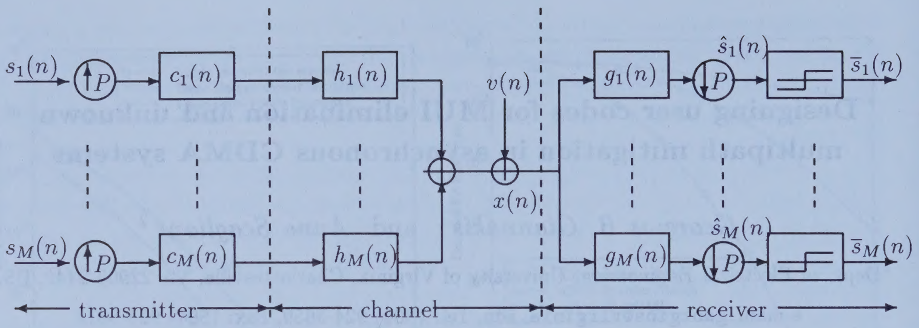


Figure 1: Multiuser/Multirate Discrete-Time Equivalent CDMA Model

2 Lagrange Codes

The block diagram in Fig. 1 represents a CDMA system, described in terms of the equivalent discrete-time baseband model, where signals, codes, and channels are represented by samples of their complex envelopes taken at the chip rate. Similar to the multirate CDMA models in [9, 10] and [12], each of the M users spreads the information bits $s_m(n)$ with the upsampler and encodes them using the code $c_m(n)$ of length P, before transmission. With $\star$ denoting convolution, the transmitted signal is:

$$u_m(n) = \sum_i s_m(i)\delta(n - iP) \star c_m(n) = \sum_i s_m(i)c_m(n - iP) \; . \tag{1}$$

We will initially assume synchronism among users and ideal propagation conditions; i.e., $h_m(n) = h_m\delta(n)$, for $m = 1, \ldots, M$, and noise $v(n) \equiv 0$. The received signal is:

$$x(n) = \sum_{m=1}^{M} \alpha_m u_m(n) = \sum_{m=1}^{M} \alpha_m \sum_i s_m(i)c_m(n - iP) \; , \tag{2}$$

where α_m denotes the received mth user's power. The mth user's information sequence can be recovered by filtering the data with the receive-filter $g_m(n)$, which must satisfy:

$$\sum_k g_m(k)c_p(n - k) = \delta(n)\delta(m - p) \; , \tag{3}$$

where $\delta(n)$ is the Kronecker delta. It is well known that in order to satisfy (3), the codes $\{c_m(n)\}_{m=1}^{M}$ must be linearly independent. The output of the filter $g_m(n)$ is then:

$$\zeta_m(n) = \sum_{p=1}^{M} \alpha_p \sum_i s_p(i) \sum_k g_m(k)c_p(n - k - iP) = \sum_i \alpha_m s_m(i)\delta(n - iP), \tag{4}$$

and the symbols can be obtained after downsampling $\zeta_m(n)$ by a factor P and removing the factor α_m:

$$s_m(n) = \zeta_m(nP)/\alpha_m. \tag{5}$$

If additive stationary noise is present, the best choice in terms of output SNR is to select codes that form an orthogonal basis and the corresponding set of receivers that are matched to the codes; i.e., $g_m(n) = c_m^*(P - n)$, where $*$ stands for conjugation.

A more realistic CDMA model however, must include asynchronism and propagation through frequency selective channels. In our model, the mth channel is characterized by an unknown impulse response $h_m(n)$, that we consider constant only during the symbol interval. The channel $h_m(n)$ can also take into account the mth user's asynchronism in the form of delay factors. Moreover, we adopt the standard conventions that: $h_m(n)$'s are FIR and the maximum order of all channels, denoted by L, is known.

The received multiplexed signal, denoted by $x(n)$, is then:

$$x(n) = \sum_{m=1}^{M} \sum_{i} s_m(i) \sum_{l} h_m(l) c_m(n - l - iP) + v(n) \ , \tag{6}$$

where $v(n)$ is additive white Gaussian noise (AWGN).

The goal is to achieve *complete MUI suppression*, irrespective of the channels' impulse responses, with low-complexity receive-filters $g_m(n)$. We summarize our assumptions as follows:

(a1)$P - M > L$, where L is the maximum expected order of all channels $\{h_m(l)\}_{m=1}^{M}$;

(a2)codes have L trailing zeros ('guard chips'); i.e., $c_m(n) = 0$ for $n = M, \ldots, P-1$;

In view of **(a1)** and **(a2)**, and considering only $n = kP, \ldots, kP+P-1$, the corresponding samples of the received signal in polyphase form are:

$$x(nP + k) = \sum_{m=1}^{M} \sum_{i} s_m(i) \sum_{l} h_m(l) c_m((n-i)P - l + k) + v(nP + k) \ , \tag{7}$$

where $k = 0, \ldots, P-1$. Taking into account the finite support of $c_m(n)$, and **(a1)**, **(a2)**, the only nonzero summand over i in (7), results when $i = n$. Thus, we can write:

$$x(nP + k) = \sum_{m=1}^{M} s_m(n) \sum_{l} h_m(l) c_m(k - l) + v(nP + k) \ , \quad k = 0, \ldots, P-1 \ . \tag{8}$$

To describe our MUI cancellation idea, it is useful to introduce a vector notation based on the following definitions:

$$\boldsymbol{x}(n) \ := \ (x(nP)x(nP+1) \cdots x(nP+P-1))^T \tag{9}$$

$$\boldsymbol{v}(n) \ := \ (v(nP)v(nP+1) \cdots v(nP+P-1))^T \tag{10}$$

$$\boldsymbol{h}_m \ := \ (h_m(0)h_m(1) \cdots h_m(L))^T \tag{11}$$

$$\boldsymbol{c}_m \ := \ (c_m(0)c_m(1) \cdots c_m(P-1))^T = (c_m(0)c_m(1) \cdots c_m(P-L-1) \ 0 \cdots 0)^T \tag{12}$$

$$\boldsymbol{g}_m \ := \ (g_m^*(P-1) \cdots g_m^*(1) g_m^*(0))^T \tag{13}$$

$$\boldsymbol{C}_m \ := \ \begin{bmatrix} c_m(0) & 0 & \ldots & 0 \\ \vdots & c_m(0) & \ddots & \\ c_m(M-1) & \vdots & \ldots & c_m(0) \\ \vdots & c_m(M-1) & \ddots & \vdots \\ 0 & \vdots & & c_m(M-1) \end{bmatrix}_{P \times (L+1)} , \tag{14}$$

where $\boldsymbol{C}_m$ is the Sylvester matrix associated with the $P \times 1$ vector $\boldsymbol{c}_m$ in (12). In view of definitions (9)-(14), eq. (8) can be written in vector form as:

$$\boldsymbol{x}(n) = \sum_{m=1}^{M} \boldsymbol{C}_m \boldsymbol{h}_m s_m(n) + \boldsymbol{v}(n) \ . \tag{15}$$

Consider without loss of generality (w.l.o.g.) the first user as the user of interest. To suppress MUI, we seek a vector $\boldsymbol{g}_1$ such that:

$$\begin{aligned} \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{x}(n) &= \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{C}_1\boldsymbol{h}_1 s_1(n) + \sum_{m=2}^{M} \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{C}_m\boldsymbol{h}_m s_m(n) + \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{v}(n) & (16) \\ &= \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{C}_1\boldsymbol{h}_1 s_1(n) + \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{v}(n)\,, & (17) \end{aligned}$$

where $^{\mathcal{H}}$ denotes transposition and conjugation. Necessary and sufficient conditions for eq. (17) to hold are:

$$\begin{aligned} \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{C}_1\boldsymbol{h}_1 &= \mathcal{K} \neq 0\,, & (18) \\ \boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{C}_m\boldsymbol{h}_m &= 0\,, \quad m \neq 1\,, & (19) \end{aligned}$$

where $\mathcal{K}$ is a constant whose magnitude depends on the transmitted power and controls the SNR at the receiver output. If the channel is known, the condition in (19) is equivalent to imposing orthogonality of $\boldsymbol{g}_1$ to all vectors $\boldsymbol{w}_m := \boldsymbol{C}_m\boldsymbol{h}_m$, for $m = 2, \ldots, M$. On the other hand, if the channel is unknown, eqs. (18) and (19) are satisfied *irrespective* of the channels' impulse responses only if:

$$\boldsymbol{g}_1^{\mathcal{H}} \in \bigcap_{m=2}^{M} \mathcal{N}(\boldsymbol{C}_m) \quad \text{and} \quad \boldsymbol{g}_1^{\mathcal{H}} \notin \mathcal{N}(\boldsymbol{C}_1)\,, \tag{20}$$

where $\mathcal{N}(\boldsymbol{C}_m)$ indicates the null space of the matrix $\boldsymbol{C}_m$. For arbitrary codes $\{c_m(n)\}_{m=1}^{M}$, the existence of a vector $\boldsymbol{g}_1$ that satisfies (20) is not guaranteed. However, judicious design of codes makes this possible and constitutes the key idea behind our work. First, we observe that the constraints in (20) can be generalized for every other user besides the first one to obtain:

$$\boldsymbol{g}_p^{\mathcal{H}} \in \mathcal{N}(\boldsymbol{C}_m) \quad \forall p \neq m; \quad p, m = 1, ..., M \quad \text{and} \quad \boldsymbol{g}_p^{\mathcal{H}} \notin \mathcal{N}(\boldsymbol{C}_p)\,. \tag{21}$$

Second, we note that due to **(a2)**, matrix $\boldsymbol{C}_m$ in (14) is full rank, and hence its nullity is $\nu(\boldsymbol{C}_m) = P - L - 1 = M - 1$. Moreover, we will prove by contradiction that vectors $\boldsymbol{g}_p^{\mathcal{H}}$ are linearly independent. Suppose w.l.o.g. that $\boldsymbol{g}_1^{\mathcal{H}}$ is a linear combination of the vectors $\{\boldsymbol{g}_p^{\mathcal{H}}\}_{p=2}^{M}$ with coefficients $(\lambda_2, \ldots, \lambda_M)$. According to (21), we have $\boldsymbol{g}_p^{\mathcal{H}} \in \mathcal{N}(\boldsymbol{C}_1)$ for $p = 2, \ldots, M$, and therefore $\boldsymbol{g}_1^{\mathcal{H}} = \sum_{p=2}^{M} \lambda_p \boldsymbol{g}_p^{\mathcal{H}} \in \mathcal{N}(\boldsymbol{C}_1)$, which contradicts (20). As a consequence, we have established the following:

Result 1: *Eq. (21) is satisfied under* **(a1)-(a2)**, *if and only if* $\forall m = 1, \ldots, M$, *the null space* $\mathcal{N}(\boldsymbol{C}_m)$ *is spanned by the* $M - 1$ *vectors* $\boldsymbol{g}_p^{\mathcal{H}}$, *with* $p \neq m$ *and* $p = 1, ..., M$.

The design criterion is a consequence of Result 1 and relies on the Sylvester structure of $\boldsymbol{C}_m$. It can be easily verified that the null vectors of a Sylvester matrix, such as $\boldsymbol{C}_m$ of (14), are the Vandermonde vectors $\boldsymbol{v} = (1\ \rho\ \rho^2\ \ldots\ \rho^{P-1})$, where ρ is a complex number. In particular, the null space of $\boldsymbol{C}_m$ is spanned by Vandermonde vectors built using as ρ's the roots of the $\mathcal{Z}$-transform of the codes $c_m(n)$. For each code of length M, we associate a polynomial $\sum_{n=0}^{M-1} c_m(n) z^{-n}$ with $M - 1$ roots $\{\rho_i\}_{i=1}^{M-1}$; i.e.,

$$\sum_{n=0}^{M-1} c_m(n) z^{-n} = \prod_{i=1}^{M-1} (1 - \rho_i z^{-1})\,. \tag{22}$$

Based on the previous observations, our system design includes the following steps:

Step 1: select M Vandermonde vectors, or, equivalently M roots $\{\rho_m\}_{m=1}^{M}$ on the complex plane;

Step 2: assign to the mth user the code associated with the $\mathcal{Z}$-domain polynomial:

$$C_m(z) = \mathcal{K}_m \frac{\prod_{i=1,i\neq m}^{M}(1-\rho_i z^{-1})}{\prod_{i=1,i\neq m}^{M}(1-\rho_i \rho_m^{-1})}\,, \qquad |\mathcal{K}_m| > 0; \tag{23}$$

Step 3: at the receiver side, assign to the mth receive-filter the Vandermonde vector built using the mth user's signature-root ρ_m:

$$\boldsymbol{g}_m^{\mathcal{H}} = (1\ \rho_m^{-1}\ \rho_m^{-2}\ \ldots\ \rho_m^{-P+1})\,. \tag{24}$$

Given a set of M complex points $\{\rho_m\}_{m=1}^{M-1}$, $\{C_m(z)\}_{m=1}^{M-1}$ in (23) is the Lagrange basis of polynomials interpolating this set of points. This is why we name our codes, Lagrange codes. Note also that $C_m(\rho_p) = \mathcal{K}_m\delta(p-m)$. The pairs $(\boldsymbol{c}_m, \boldsymbol{g}_m^{\mathcal{H}})$ satisfy eq. (21), since:

$$\boldsymbol{g}_p^{\mathcal{H}}\boldsymbol{C}_m = C_m(\rho_p)(1\ \rho_p^{-1}\ \rho_p^{-2}\ldots\rho_p^{-L}) = \mathcal{K}_p\delta(m-p)(1\ \rho_p^{-1}\ \rho_p^{-2}\ldots\rho_p^{-L})\,. \tag{25}$$

From (25), it follows that the MUI can be simplified, and (17) becomes:

$$\boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{x}(n) = \mathcal{K}_1 H_1(\rho_1)s_1(n)\ +\boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{v}(n)\,, \tag{26}$$

where $H_m(\rho) := \sum_{l=0}^{L} h_m(l)\rho^{-l} = (1\rho_p^{-1}\rho_p^{-2}\ldots\rho_p^{-L})\boldsymbol{h}_m$ and $\boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{v}(n) = \sum_{k=0}^{P-1} v(nP+k)\rho_1^{-k}$. The same argument applies to all other users.

Orthogonal frequency division multiple access (OFDMA) is subsumed in our design as we summarize next:

Result 2: *OFDMA follows as a special case of the Lagrange codes, if one selects* $\rho_m = \exp(j2\pi m/M)$. *The coefficients of* $C_m(z)$ *in (23) are (within a scale) given by* $c_m(n) = \exp(j2\pi mn/M)$.

3 Performance and Simulations

The transmission strategy based on Lagrange codes offers advantages relative to existing blind approaches, because the receiver is precomputable and achieves without extra computations the same goal as in [5] and [8]. As far as MUI suppression in multipath is concerned, minimal assumptions are invoked because we only require that the channel variation is negligible during the symbol interval. In this section, we will evaluate the performance of the proposed transmission strategy for linear time invariant (LTI) as well as for rapidly time varying channels induced by Doppler frequency shifts.

3.1 LTI Multipath Channels

From (26), we observe that the effects of multipath propagation and noise are present in the form of the unknown factor $H_1(\rho_1)$ ($H_m(\rho_m)$ for the mth user) and the noise term $\boldsymbol{g}_1^{\mathcal{H}}\boldsymbol{v}(n)$ ($\boldsymbol{g}_m^{\mathcal{H}}\boldsymbol{v}(n)$).

To compensate for the unknown multiplicative factor $H_m(\rho_m)$, differential encoding and demodulation of the transmitted symbols $s_m(n)$ can be adopted. Performance analysis for general Qth order PSK constellations is then possible by computing error probabilities for coherent and non coherent demodulation of differentially encoded PSK symbols (see e.g., [1]). Probability of error for the mth user depends on the signal-to-noise ratio (SNR_m) at the output of the $g_m(n)$ filter. Since $\boldsymbol{v}(n)$ is an AWGN vector the noise term in (26), $\boldsymbol{g}_m^{\mathcal{H}}\boldsymbol{v}(n)$, is a Gaussian random variable, with zero mean and variance:

$$E\{|\boldsymbol{g}_m^{\mathcal{H}}\boldsymbol{v}(n)|^2\} = \sigma_v^2\boldsymbol{g}_m^{\mathcal{H}}\boldsymbol{g}_m = \sigma_v^2\frac{1-|\rho_m|^{-2P}}{1-|\rho_m|^{-2}} . \tag{27}$$

Based on (26) and (27), SNR_m is given by:

$$SNR_m := \frac{|\boldsymbol{g}_m^{\mathcal{H}}\boldsymbol{C}_m\boldsymbol{h}_m|^2}{\sigma_v^2\boldsymbol{g}_m^{\mathcal{H}}\boldsymbol{g}_m} = \frac{|\mathcal{K}_m|^2|H_m(\rho_m)|^2(1-|\rho_m|^{-2})}{\sigma_v^2(1-|\rho_m|^{-2P})} . \tag{28}$$

The factor $(1-|\rho_m|^{-2})/(1-|\rho_m|^{-2P})$ in (28) is a monotonically increasing function of $|\rho_m|$. Based on this factor alone, one might be tempted to maximize the SNR_m by increasing $|\rho_m|$. However, we will prove that this factor cannot be made arbitrarily large without requiring unbounded transmitted energy. Towards this goal, let us define the energy per symbol E_s as: $E_s := \boldsymbol{c}_m^{\mathcal{H}}\boldsymbol{c}_m$, and suppose w.l.o.g. that it is the same for every user. Using Parseval's relation, we can express E_s in the $\mathcal{Z}$-domain as:

$$E_s = |\mathcal{K}_m|^2\frac{1}{2\pi j}\oint\prod_{i=1,i\neq m}^{M-1}\frac{(1-\rho_i z^{-1})(1-\rho_i^* z)}{|1-\rho_i\rho_m^{-1}|^2}z^{-1}dz . \tag{29}$$

If there exists an $i \neq m$ such that $|\rho_i| \to \infty$, the right hand side of (29) goes to infinity, implying that $|\mathcal{K}_m|^2 \to 0$, for E_s finite. The latter means that if one chooses to maximize the SNR_i corresponding to the ith user by setting its signature root $|\rho_i| = \infty$, then all the other users' codes must be necessarily annihilated, if E_s is to be finite.

Having established that $|\rho_m|$ must be finite, maximization of SNR_m for fixed $|\rho_m|$ values suggests that the optimal receive-filters should be *matched* to the corresponding codes; i.e., $\boldsymbol{g}_m^{\mathcal{H}} = \boldsymbol{c}_m^T$. But (21) requires that $\boldsymbol{g}_m^{\mathcal{H}}\boldsymbol{c}_p = \mathcal{K}_m\delta(m-p)$, and since $\boldsymbol{g}_m^{\mathcal{H}} = \boldsymbol{c}_m^T$, this is equivalent to requiring that the codes $\boldsymbol{c}_m$ (and hence $\boldsymbol{g}_m$) are orthogonal vectors. On the other hand, to assure that MUI is suppressed, the receive-filters have to be Vandermonde vectors. It is easy to verify that the only *orthogonal* Vandermonde vectors are $c_m(n) = g_m^*(P-n) = \mathcal{K}_m\exp(j2\pi mn/M)$, which are the complex exponentials used in OFDMA. We have thus established that:

Result 3: *Under the finite transmitted energy constraint, Lagrange codes with roots uniformly spaced around the unit circle (corresponding to OFDMA transmission) maximize the SNR_m for non-frequency selective channels; i.e., when $|H_m(\rho_m)|^2 = 1$.*

Example 1: (Output SNR as a function of code roots) Fig. 2 illustrates how SNR_m is influenced by the locations of the code-roots when our Lagrange transmitter's energy is fixed. All $M = 16$ users transmitted their bits with code roots $\rho_m = R\exp(j2\pi m/M)$, equispaced on a circle of radius R at a fixed energy $E_s = 1$. Due to the symmetric

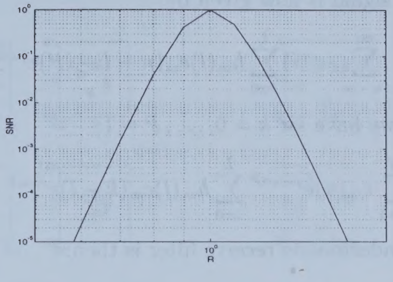
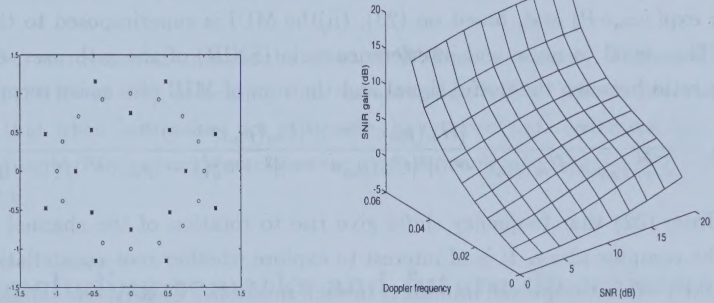


Figure 2: Lagrange code-roots $\rho_m = R\exp(j2\pi m/M)$, $E_s = 1$, $M = 1$.


(a)

(b)

Figure 3: (a) Roots of user codes; (b) SNIR gain (dB) of Lagrange codes vs. OFDMA.

location of the roots, the output SNR for all users is characterized by the same factor $|\mathcal{K}_m|^2(1 - |\rho_m|^{-2})/\sigma_v^2(1 - |\rho_m|^{-2P})$. It is evident that moving the roots outside or inside the unit circle, decreases the ratio $|\mathcal{K}_m|^2(1 - |\rho_m|^{-2})/\sigma_v^2(1 - |\rho_m|^{-2P})$.

Remark: If the channel impulse responses are known at the transmitter side, the location of the roots may be optimized to maximize the SNR_m along the lines of [7]. This involves a multivariate nonlinear optimization procedure and goes beyond the scope of the present paper, but constitutes an interesting direction for future research.

3.2 Time-Varying Multipath Channels

The freedom in designing Lagrange codes can also be exploited to reduce sensitivity to carrier frequency shifts, due to Doppler effects, or, imperfect carrier synchronization between transmitter and receiver. The presence of a frequency shift impairs MUI suppression capabilities, and is modeled as a multiplicative effect on the output samples of the mth linear channel, $\{h_m(l)\}_{l=0}^{L}$, via the unknown complex exponential $\exp(j\omega_m n)$.

The received multiplexed signal is now given by:

$$x(n) = \sum_{m=1}^{M} \sum_{i=-\infty}^{\infty} s_m(i) \sum_{l=0}^{L} h_m(l) c_m(n-l-iP) e^{j\omega_m n} + v(n). \tag{30}$$

Proceeding as in eq. (8), we have for $k = 0, \ldots, P-1$:

$$x(nP+k) = \sum_{m=1}^{M} s_m(n) e^{j\omega_m nP} \sum_{l=0}^{L} h_m(l) c_m(k-l) e^{j\omega_m k} + v(nP+k)\,. \tag{31}$$

The output of the pth Vandermonde receive-filter is then:

$$\begin{aligned} \boldsymbol{g}_p^{\mathcal{H}} \boldsymbol{x}(n) &= \sum_{k=0}^{P-1} x(nP+k) \rho_p^{-k} \\ &= \sum_{m=1}^{M} s_m(n) e^{j\omega_m nP} H_m(\rho_p e^{-j\omega_m}) C_m(\rho_p e^{-j\omega_m}) + \sum_{k=0}^{P-1} v(nP+k) \rho_p^{-k}\,. \end{aligned} \tag{32}$$

From (32), we observe that two undesired effects arise: (i)the symbols are modulated by the factor $\exp(j\omega_m nP)$ and, based on (23); (ii)the MUI is superimposed to the user of interest. The signal to noise and interference ratio (SNIR) of the mth user, defined as the power ratio between the useful signal and the sum of MUI plus noise term, is:

$$SNIR_m = \frac{|H_m(\rho_m e^{-j\omega_m})|^2 |C_m(\rho_m e^{-j\omega_m})|^2}{\sum_{p=1, p\neq m}^{M} |H_m(\rho_m e^{-j\omega_p})|^2 |C_m(\rho_m e^{-j\omega_p})|^2 + \sigma_v^2 (1 - |\rho_m|^{-2P})/(1 - |\rho_m|^{-2})}\,. \tag{33}$$

We note from (32) that frequency shifts give rise to rotation of the channel and code roots in the complex plane. It is of interest to explore whether root constellations other than OFDMA exhibit improved immunity to such rotations. Clearly, one should increase the Euclidean distance among roots in order to reduce the possibility of mismatching signature roots when they rotate relative to the roots used at the Vandermonde receiver.

Example 2: (Lagrange vs OFDMA codes with frequency drifts) Here we simulated $M = 16$ users with code roots, $\{\rho_i = 1.1 \exp(j\pi i/M)\}_{i=1}^{7}$ and $\{\rho_i = 0.9 \exp(j\pi(i+0.5)/M)\}_{i=8}^{16}$. The root constellation is depicted in Fig. 3a with stars along with the OFDMA configuration shown with circles (note the wider angle between roots located at the same distance from the origin, relative to OFDMA). The comparison between the two configurations was carried for different SNR and frequency shift offset values $\omega_D/2\pi$. Fig. 3b shows their average SNIR ratio (referred to as SNIR gain) vs. the frequency shift ($\omega_m = \omega_D \ \forall\, m$) and the SNR, under ideal channels $\{H_m(\rho)\}_{m=1}^{M} = 1, \forall\, m$. It is evident that unless the frequency shift is very small, the proposed scheme clearly outperforms OFDMA and the gain increases as the frequency shift increases.

Example 3: (Frequency-selective multipath channels) Here we simulated FIR channels of length $L = 4$, with first coefficient one and the other coefficients selected as independent complex Gaussian random variables with zero mean and standard deviation 0.7. The frequency shifts were assumed to be $\omega_D = 2\pi/120$, the same for all channels. Fig. 4 shows the scattering diagrams for the Vandermonde receiver outputs of the first 9 (out of 16) receive-filters based on OFDMA and the proposed Lagrange codes (the factor

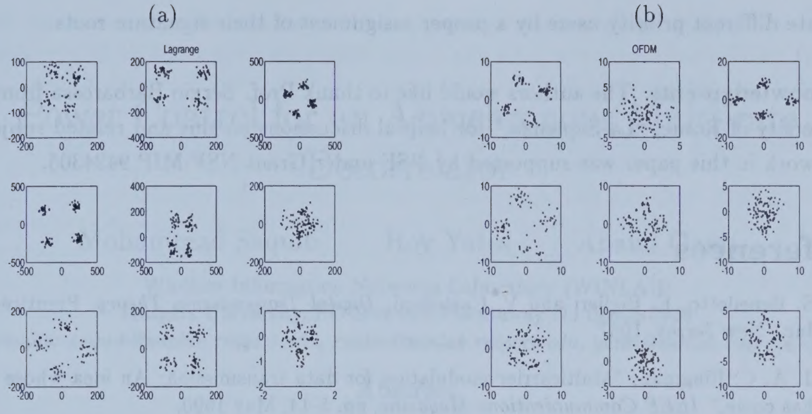


Figure 4: Eye diagrams of (a)Lagrange; (b)OFDMA ($L = 4$, $M = 16$, $SNR = 40dB$).

$\exp(j\omega_m Pn)$ was removed to make the effect of MUI more evident). The noise level was very small (SNR=40 dB), indicating that the spreading in Fig. 4 is mainly due to MUI. It is clear that when fading does not attenuate the received data considerably, the clouds corresponding to Vandermonde receivers are more concentrated than those corresponding to OFDMA.

4 Concluding remarks and future directions

We introduced a novel code-design methodology for low-complexity CDMA receivers capable of suppressing MUI completely in the presence of *unknown* and even rapidly varying multipath. Contrary to existing blind approaches, the proposed Vandermonde receivers do not rely on the data because they can be pre-computed from appropriately designed Lagrange codes.

The code polynomials were considered to have fixed roots. However, if the channel is available, we can allocate the roots among users optimally. Specifically, if the only available information about the channel is its order L, recalling (26), we can reduce the probability that the mth code would be suppressed by the channel by assigning more than one signature root to each user. This entails the use of longer codes, and hence extra redundancy. In principle, assigning $L + 1$ roots to each user, which means using codes of length $(L+1)M$, it is possible to assure that the mth user cannot be annihilated by the channel. If the channels are known, we can also optimize the locations of the roots, without introducing any further redundancy, as suggested in Section 3.

Even when all the users have the same energy, it is important to observe from (28) that for an arbitrary root assignment the performance in terms of SNR_m depends on the particular signature root associated to the mth user. This offers freedom to accommodate different priority users by a proper assignment of their signature roots.

Acknowledgments: The authors would like to thank Prof. Sergio Barbarossa from the University of Rome, "La Sapienza," for helpful discussions on this and related subjects. The work in this paper was supported by NSF under Grant NSF-MIP 9424305.

References

[1] S. Benedetto, E. Biglieri and V. Castellani, *Digital Transmission Theory,* Prentice-Hall Inc., New Jersey, 1987.

[2] J. A. C. Bingman, "Multicarrier modulation for data transmission: An idea whose time has come," *IEEE Communications Magazine*, pp. 5-14, May 1990.

[3] L. J. Cimini, "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiple access," *IEEE Trans. on Communications*, pp. 665-675, 1995.

[4] M. L. Honig, U. Madhow, and S. Verdú, "Blind adaptive multiuser detection," *IEEE Trans. on Info. Theory*, pp. 944–96, July 1995.

[5] H. Liu and G. Xu, "A subspace method for signature waveform estimation in synchronous CDMA systems,"*IEEE Trans. on Communications*, pp. 1346–1354, November 1996.

[6] R. Lupas and S. Verdu, "Near-far resistance of multiuser detectors in asynchronous channels." *IEEE Trans. on Comm.*, pp. 496–508, Apr. 1990.

[7] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Redundant Filterbank Precoders and Equalizers: Optimal Designs and Blind Signal Recovery," *Proc. of 35th Allerton Conf. on Comm., Control, & Comp.*, September 1997.

[8] M. K. Tsatsanis, "Inverse filtering criteria for CDMA systems", *IEEE Trans. on Signal Processing*, pp. 102–112, Jan. 1997.

[9] M. K. Tsatsanis and G. B. Giannakis, "Multirate filter banks for code-division multiple access systems," *Proc. of Intl. Conf. on ASSP,* vol. 2, pp. 1484-1487, 1995.

[10] M. K. Tsatsanis and G. B. Giannakis, "Optimal linear receivers for DS-CDMA systems: a signal processing approach," *IEEE Trans. on Signal Processing,* pp. 3044-3055, 1996.

[11] S. Verdu, "Multiuser Detection," in *Advances in Statistical Signal Processing*, V. Poor Edit., NY, JAI, vol. 2, pp. 369-409, 1993.

[12] G.Wornell, "Emerging applications of multirate signal processing and wavelets in digital communications," *Proceedings of the IEEE*, pp. 586-603, April 1996.

Power Control for an Asynchronous Multi-rate Decorrelator

Mohammad Saquib Roy Yates Anand Ganti

Wireless Information Networks Laboratory (WINLAB)
Rutgers University, PO Box 909 Piscataway NJ 08855-0909
e-mails: *saquib@winlab.rutgers.edu*, *ryates@winlab.rutgers.edu*, *ganti@winlab.rutgers.edu*

Abstract

The varied Quality of Service (QoS) or bit error rate (BER) requirements of multimedia traffic require the use of power control for CDMA wireless systems employing multiuser detection. Using a decorrelator in an asynchronous multi-rate DS/CDMA system, it may be necessary for different users to combat the noise enhancement and the propagation losses to varying degrees depending on individual requirements. In this context, we propose a power control algorithm for a multi-rate decorrelator that is suitable for a class of BER based link quality objectives. If the uplink channel gain of the desired user is known, then it is simple for each user to choose the transmitter power needed to meet its target BER objective. In practice, however, the uplink channel gain is often difficult to measure. To avoid this measurement, we employ stochastic approximation methods to develop a simple, distributed, iterative power control algorithm. In this algorithm, each mobile will use the output of its own decorrelator to update its transmitted power in order to achieve its QoS objective. We will show that when a user's bits have nonzero asymptotic efficiencies, then the power control algorithm converges quickly in the mean square sense to an optimal power at which the desired user achieves its QoS objective.

1 Introduction

Extensive efforts are being made to integrate code division multiple access (CDMA) cellular networks with fixed networks for communicating both voice and data messages. Receiver design for an asynchronous multi-rate CDMA system has become a new area of investigation [1-4]. Reference [2] proposed a decorrelator called the *asynchronous multi-rate decorrelator* or *AMD* for an asynchronous multi-rate CDMA system. In the asynchronous multi-rate system, variable processing gains were used to implement distinct bit rates. The AMD is decentralized and decodes users by employing sliding finite length observation windows. It was proven in [2] that the performance of any finite window decorrelator improves monotonically with the observation window length, although a longer observation window results in greater computational complexity. The AMD operates by employing a set of modified correlators whose signals are simply time shifted versions of one another and can decode a user by suppressing the multiuser interference as long as the desired user has non-zero asymptotic efficiency [5]. As with all decorrelators, the multiuser interference is suppressed at the cost of reduced received signal energy. As a consequence of supporting multiple rates in an asynchronous system, different bits of an individual user are subject to varying degrees of energy degradation and, as a consequence, may experience different probabilities of error.

Recently, several studies [6,7] have been performed in order to integrate power control with multiuser detectors. The motivation of these works was to achieve a performance gain over multiuser detection by providing power control for multiuser detection. Reference [6] addressed power control for the linear MMSE multiuser receiver in a single-rate system. The proposed schemes are synchronous and require SIR (signal-to-interference ratio) measurements for the power updates.

In this paper, a class of QoS objectives for the AMD are proposed. The choice of a QoS objective will dictate a corresponding power control algorithm. If the uplink channel gain of the mobile is known then it is simple to provide each user its QoS objective. In practice, however, the measurement of the channel gain is non-trivial and neglects the stochastic behavior of the channel. This practical situation leads us to use stochastic approximation methods [8] to develop an asynchronous, distributed, iterative power control algorithm for the AMD that finds the correct powers without knowledge of the uplink gain. In this stochastic power control algorithm, a mobile needs to know only the output of the AMD to update its transmitted power in order to achieve its QoS objective. The proposed power control algorithm is feasible as long as asymptotic efficiencies of the bits of the desired user are nonzero. In this case, the algorithm converges in the mean square sense to an optimal power at which the user achieves its QoS requirement.

2 Asynchronous multi-rate system

In the asynchronous multi-rate DS/CDMA system model of [2], each bit results in the baseband transmission of a sequence of pulses, or chips, $p[t]$, each pulse having a duration of one chip period T_c. These pulses are sent over an additive white Gaussian noise channel in which the noise $N(t)$ has power spectral density σ^2.

For a system with K users, the transmission rate of user j is denoted by $R_j = M_j R$ for an integer $M_j \geq 1$ where R denotes a system base rate. The bit transmission time of user j is $T_j = 1/R_j$ and the processing gain is $L_j = T_j/T_c$. The jth user has signature waveform $S_j(t) = \sum_{m=1}^{L_j}\{a_j(m)\frac{1}{\sqrt{L_j}}p[t-(m-1)T_c]\}$, where $t \in [0, T_j]$ and $a_j(m) \in \{-1, 1\}$ denotes the signature sequence of user j. The energy of the pulse $p[t]$ is normalized so that each $S_j(t)$ has unit energy over $[0, T_j]$. Let us denote Δ_j as the delay of the jth user. In the asynchronous channel, the received signal due to the jth user is given by

$$r_j(t) = \sum_{i=-\infty}^{+\infty} b_j^{(i)} \sqrt{E_j^{(i)}} S_j(t - iT_j - \Delta_j) \qquad 1 \leq j \leq K \tag{1}$$

where $b_j^{(i)} \in \{-1, +1\}$ is the ith bit and $E_j^{(i)}$ is the received energy of the jth user. Since the transmitted power is proportional to the received energy, we will refer to $E_j^{(i)}$ as power. It is assumed that the receiver knows the time delay Δ_j for each user j and the received signal is

$$r(t) = \sum_{k=1}^{K} r_k(t) + N(t) \tag{2}$$

We now describe how the ith bit $b_j^{(i)}$, $0 \leq i \leq M_j - 1$, of user j is decoded. In order to decode $b_j^{(i)}$, we assume $\Delta_j = 0$ and that the other Δ_k are computed relative to user j. It may be desirable for different bits to be decorrelated using different window sizes. The bit $b_j^{(i)}$ will be decoded by processing the windowed received signal

$$R_j^{(i)}(t) = r(t)\left[u\left(t - [i - n_j^{(i)}]T_j\right) - u\left(t - [i + 1 + d_j^{(i)}]T_j\right)\right] \tag{3}$$

where $u(t)$ is the unit step function. With respect to bit $b_j^{(i)}$, the window of $R_j^{(i)}(t)$ extends $n_j^{(i)}$ bits into the past and $d_j^{(i)}$ bits into the future and thus covers $n_j^{(i)} + d_j^{(i)} + 1$ bits of user j.

3 Asynchronous multi-rate decorrelator, AMD

The decorrelator [5] can be implemented in a decentralized fashion using modified correlator signals. In [2], it was mentioned that the modified correlators of a decorrelator can be generated by applying the Gram-Schmidt orthogonalization to the interfering users' signature sequences. Reference [2] focused its attention on the Gram-Schmidt procedure to develop the AMD which we now briefly describe.

Among the M_j bits of user j, we present the AMD decoding technique for bit $b_j^{(i)}$. Let $I_{n_j^{(i)}, d_j^{(i)}}$ denote the set of signature waveforms that are generated by the interfering users within the observation window $\left[(i - n_j^{(i)})T_j, (i + 1 + d_j^{(i)})T_j\right]$. The AMD decodes the bit $b_j^{(i)}$ using a modified correlator $\Phi_0^{(i)}(t)$ such that $\Phi_0^{(i)}(t)$ is orthogonal to each interfering signal in the set $I_{n_j^{(i)}, d_j^{(i)}}$. The output of the modified correlator $\Phi_0^{(i)}(t)$ is

$$r_j^{(i)} = \int_{(i-n_j^{(i)})T_j}^{(i+1+d_j^{(i)})T_j} R_j^{(i)}(t)\Phi_0^{(i)}(t)\,dt = \sqrt{\zeta_j^{(i)} E_j^{(i)}}\, b_j^{(i)} + N_j^{(i)} \tag{4}$$

where $N_j^{(i)}$ is a Gaussian random variable with mean zero and variance σ^2 and

$$\zeta_j^{(i)} = \left\{ \int_{(i-n_j^{(i)})T_j}^{(i+1+d_j^{(i)})T_j} \Phi_0^{(i)}(t) S_j(t - iT_j)\,dt \right\}^2 \tag{5}$$

is the asymptotic efficiency of the AMD for decoding the bit $b_j^{(i)}$ in the observation window $\left[(i - n_j^{(i)})T_j, (i + 1 + d_j^{(i)})T_j\right]$. The corresponding decoding rule will be $\bar{b}_j^{(i)} = \mathrm{sgn}(r_j^{(i)})$ and the probability that the bit $b_j^{(i)}$ is decoded incorrectly is

$$q_j^{(i)} = Q\left(\sqrt{\zeta_j^{(i)} E_j^{(i)} / \sigma^2}\right) \tag{6}$$

where $Q(\cdot)$ is the standard normal complementary CDF.

We have described the decoding technique for the ith bit, $i = 1, \ldots, M_j - 1$, of user j. After an interval of M_j bits, user j will observe the same pattern of interfering signature waveforms. Thus, the receiver dedicated to the user j needs to know the specifications of M_j modified correlators. Furthermore, user j will experience M_j different asymptotic efficiencies or BERs.

4 Quality of Service for the AMD

In this section, we will propose a class of QoS objectives for the AMD. Let us denote the uplink channel gain of user j by h_j and its transmitted power in the ith slot by $P_j^{(i)}$. Then the received power at the ith slot becomes $E_j^{(i)} = h_j P_j^{(i)}$. The criterion that is used for power control is to guarantee each user's link quality objective or QoS requirement. We will consider the following link quality objectives.

- *The fixed BER (FB) objective:* Provide each bit of user j with a target probability of error q_j^*. Mathematically, we must choose a set of transmitter powers $P_j^{(i)}$, $i = i, \ldots, M_j - 1$, such that

$$Q\left(\sqrt{\zeta_j^{(i)} h_j P_j^{(i)} / \sigma^2}\right) = q_j^* \qquad i = 0, \ldots, M_j - 1 \tag{7}$$

Note that the FB objective requires that the transmitter power of user j be adjusted from bit to bit. When the target BER q_j^* is low, then small variations in $\zeta_j^{(i)}$ will significantly change the transmitted power. Thus it may not be possible for user j to adjust the transmitter power quickly enough. This situation leads us to propose the following link quality objective.

- *Average BER (AB) objective:* Using a fixed transmitted power, guarantee user j an average bit error target q_j^* over M_j transmitted bits. Here $P_j^{(i)} = P_j$ for each bit $b_j^{(i)}$ and the link quality objective of user j is

$$\frac{1}{M_j} \sum_{i=0}^{M_j-1} Q\left(\sqrt{\zeta_j^{(i)} h_j P_j / \sigma^2}\right) = q_j^* \tag{8}$$

We observe that the AB objective achieves the same average BER as the FB objective by the use of a fixed transmitter power. The AB objective yields different BERs for different bits. The bit with the lowest asymptotic efficiency will experience the highest BER. In an asynchronous multi-rate channel, several bits of user j may experience significantly lower asymptotic efficiencies than others. In this case, user j may prefer to protect those bits which are suffering from low asymptotic efficiencies. To handle this situation, we propose the following link quality objective.

- *Worst case BER (WB) objective:* Using a fixed transmitted power, provide target BER q_j^* to the bit with the lowest asymptotic efficiency. That is, we choose P_j such that

$$\max_i Q\left(\sqrt{\zeta_j^{(i)} h_j P_j / \sigma^2}\right) = q_j^* \tag{9}$$

Compared to the AB objective, the WB objective will demand higher transmitted power from the mobile. We will see that the same power control algorithm can be used to achieve either the AB or WB objective. For either objective, user j will have a target received power E_j^* that guarantees the corresponding link QoS. Of course, E_j^* will depend on whether user j desires the AB or WB objective. Since $Q(\cdot)$ is a monotone decreasing function, a straightforward offline calculation can determine E_j^* for either objective. If the uplink channel gain h_j is known, then the target transmitted power P_j^* is simply E_j^*/h_j.

In practice, the uplink channel gain is often difficult to measure. The base station may require training sequences from the mobile in order to measure h_j. We develop a stochastic approximation method [8] that uses the AMD outputs to iteratively converge to a target received power E_j^* without the knowledge of the uplink channel gain h_j. Let us refer this stochastic approximation method as the *target power* or *TP* algorithm. Based on the value of E_j^*, the TP algorithm will achieve the AB or WB objective. In the TP algorithm, the base station will use the modified correlator outputs to estimate

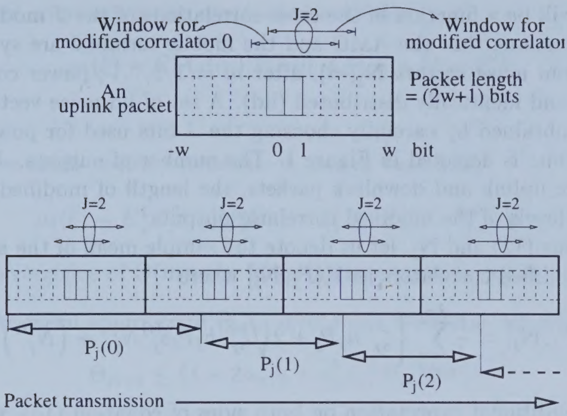


Figure 1: The modified correlators whose outputs are used in different power control iterations operate over different time intervals. Therefore, random noise vectors $\mathrm{N_j}(m)$, $m = 1, 2, \ldots$, are identically distributed and statistically independent.

the average received power and the mobile uses this estimated average received power to update its transmitted power in order to meet the link quality objective. It is easy to verify that the TP algorithm is feasible as long as the asymptotic efficiencies of user j are nonzero.

In the TP algorithm, we assume that both in the uplink (from the mobile to the base) and the downlink (from the base to the mobile), bits are transmitted in packets. It is also assumed that the base station measures J modified correlator outputs and estimates the average received power over J bits. Let us concentrate on the J modified correlators which are used to decode the bits numbered from 0 through $J-1$. From equation (4), the output for the ith bit of user j is $r_j^{(i)}$ then

$$r_j^{(i)} = \sqrt{\zeta_j^{(i)} h_j P_j} b_j^{(i)} + N_j^{(i)} \tag{10}$$

Under the transmitted power P_j, the power of user j, averaged over the outputs of J modified correlators of user j, is

$$\overline{Y}_j(P_j) = \frac{h_j P_j}{J} \sum_{i=0}^{J-1} \zeta_j^{(i)} + \sigma^2 \tag{11}$$

In equation (11), we see that $P_j = P_j^*$ iff $\overline{Y}_j(P_j) = \overline{Y}_j(P_j^*) = Y_j^*$. Therefore, the target power P_j^* will be achieved if the average power at the output of modified correlators equals Y_j^*. In the TP algorithm, the base station will transmit the average of the squared values of the J modified correlator outputs to user j who will use this value to update its transmitted power in order to achieve the target average power, Y_j^*.

Squaring equation (10), we get

$$\left(r_j^{(i)}\right)^2 = \zeta_j^{(i)} h_j P_j + 2\sqrt{\zeta_j^{(i)} h_j P_j} b_j^{(i)} N_j^{(i)} + \left(N_j^{(i)}\right)^2 \tag{12}$$

The noise components at the J modified correlator outputs can be represented by the noise vector $\mathrm{N_j} = \left[N_j^{(0)}, \ldots, N_j^{(J-1)}\right]$. Note that the autocorrelation matrix of the noise vector $\mathbf{N_j}$ will be a function of the cross-correlations of the J modified correlators. We will need to assume that the AMD and the mobile terminal are synchronized such a way that random noise vectors $\mathbf{N_j}(m)$, after $m = 1, 2, \ldots$, power control iterations, are independent and identically distributed (iid). A set of iid noise vectors $\mathbf{N_j}(m)$, $m = 1, 2, \ldots$, can be obtained by carefully choosing the J bits used for power control. One such choice of J bits is depicted in Figure 1. The number of outputs, J will depend on the lengths of the uplink and downlink packets, the length of modified correlators and the quantization levels of the modified correlator outputs.

Using equations (12) and $\mathbf{N_j}$, let us denote the sample mean of the average power at the outputs of modified correlators as $Y_j[P_j, \mathbf{N_j}]$ where

$$Y_j[P_j, \mathbf{N_j}] = \frac{1}{J} \sum_{i=0}^{J-1} \left\{ \zeta_j^{(i)} h_j P_j + 2\sqrt{\zeta_j^{(i)} h_j P_j} b_j^{(i)} N_j^{(i)} + \left(N_j^{(i)} \right)^2 \right\} \tag{13}$$

Taking the conditional expectation on both sides of equation (13), we can write the average power at the outputs of modified correlators as

$$E[Y_j[P_j, \mathbf{N_j}] | P_j = \xi] = \frac{h_j \xi}{J} \sum_{i=0}^{J-1} \zeta_j^{(i)} + \sigma^2 = \overline{Y}_j(\xi) \tag{14}$$

When $\xi = P_j^*$, we obtain the target average power, Y_j^* at the outputs of modified correlators. We have mentioned earlier that in the TP algorithm, our goal is to obtain $\overline{Y}_j(\xi) = Y_j^*$ to guarantee user j the target power P_j^*. Note that exact calculation of Y_j^* requires perfect estimation of the noise power or variance, σ^2. Using $Y_j[P_j, \mathbf{N_j}]$ and Y_j^*, the TP algorithm for user j implements the stochastic approximation method

$$P_j(m+1) = P_j(m) - a_m \left(Y_j[P_j(m), \mathbf{N_j}(m)] - Y_j^* \right) \tag{15}$$

Note that a_m may be a constant or a function of the iteration index m. The properties of a_m that are necessary for the convergence of the proposed equation (15) will be discussed shortly. Note that either AB or WB objectives can be achieved by appropriately setting the target average received power Y_j^*. The mobile executes the TP algorithm without knowledge of the asymptotic efficiencies $\zeta_j^{(i)}$.

5 Convergence of the TP algorithm

Let us denote the *mean square error* of the transmitter power at the mth iteration by

$$\Theta_m = E\Big[\big[P_j(m) - P_j^* \big]^2 \Big] \tag{16}$$

Our goal will be to verify that the TP algorithm converges by showing that

$$\lim_{m \to \infty} \Theta_m = 0 \tag{17}$$

The evolution of the stochastic power control algorithm is dictated by the received power error $X_j(m) = Y_j[P_j(m), \mathbf{N_j}(m)] - Y_j^*$. We will use upper bounds on the first and second conditional moments of $X_j(m)$ given $P_j(m)$ in order to prove convergence. Using $\mu_k(\xi) = E\big[X_j^k(m) | P_j(m) = \xi\big]$ to denote the kth conditional moment, we will need the following lemmas. Due to space constraints, proofs will be found in [9].

Lemma 1 *There exists a positive constant* c_0 *such that*

$$\mu_1(\xi) = E[X_j(m)|P_j(m) = \xi] = c_0\left(\xi - P_j^*\right) \tag{18}$$

Lemma 2 *There exists positive constants* c_1 *and* c_2 *such that*

$$\mu_2(\xi) = E\big[X_j^2(m)|P_j(m) = \xi\big] \leq c_1\left(\xi - P_j^*\right)^2 + c_2 \tag{19}$$

To prove the convergence of $\Theta(m)$, the following intermediate result will be used.

Theorem 1 *The mean square error* Θ_{m+1} *after* $m+1$ *iterations satisfies the upper bound*

$$\Theta_{m+1} \leq \left(1 - 2a_m c_0 + a_m^2 c_1\right)\Theta_m + a_m^2 c_2 \tag{20}$$

Theorem 1 implies that the limit (17) will hold if

$$\lim_{m\to\infty}\left[\left(1 - 2a_m c_0 + a_m^2 c_1\right)\Theta_m + a_m^2 c_2\right] = 0 \tag{21}$$

In this study, we will establish the limit (17) by proving the convergence of (21).

To study the convergence of the TP algorithm for a fixed a_m, we prove the following theorem.

Theorem 2 *If* $a_m = a$ *and* $|1 - 2ac_0 + a^2c_1| \leq 1$, *then*

$$\lim_{m\to\infty}\Theta_m \leq \frac{ac_2}{2c_0 - ac_1} \tag{22}$$

Note that $|1 - 2ac_0 + a^2c_1| \leq 1$ implies $2c_0 \geq ac_1$ and thus, the upper bound of Theorem 2 is always positive. Theorem 2 implies that choosing an arbitrarily small a, we can obtain $\frac{ac_2}{2c_0 - ac_1} \approx 0$. However, as a approaches zero, $|1 - 2ac_0 + a^2c_1|$ approaches 1 and the convergence speed slows down. This problem also arises when c_0 and c_1 are very small. In this case, we need to choose a very large a to achieve $|1 - 2ac_0 + a^2c_1| < 1$. Suitable selection of a requires the knowledge of c_0, c_1 and c_2. As c_0 and c_1 are functions of the uplink channel gain h_j, thus, the selection of a is as difficult as the measurement of h_j.

This practical situation leads us to search a coefficient sequence a_m which will always guarantee $\lim_{m\to\infty}\Theta_m = 0$. In the following theorem, we show that if a_m satisfies two simple conditions, then $\lim_{m\to\infty}\Theta_m = 0$.

Theorem 3 *If the sequence* a_m *satisfies*

$$\sum_{m=1}^{\infty} a_m = \infty \qquad \sum_{m=1}^{\infty} a_m^2 < \infty$$

then $\lim_{m\to\infty}\Theta_m = \lim_{m\to\infty} E\Big[\left(P_j(m) - P_j^*\right)^2\Big] = 0.$

Note that $a_m = a/m$ meets both conditions of Theorem 3. The proof of Theorem 3 follows Sakrison's procedure [8, pages 60-61].

Although the TP algorithm will converge, a number of issues must be resolved in order to implement the algorithm. In practice, the uplink channel gain h_j and the noise variance σ^2 are very small. Therefore, the received power measurement $Y_j[P_j, \mathrm{N_j}(m)]$

and the target received power Y_j^* in equation (15) will be very small. The TP algorithm uses these small values to update the transmitted power which typically is several orders of magnitude larger. If the initial transmitted power varies significantly from P_j^*, then the convergence speed of the TP algorithm will be extremely slow. For fast convergence, we need to choose a sufficiently large. The heuristic procedure for the selection of a will be discussed in the following section.

A second practical concern is that the channel gain h_j, the time offsets of equation (1), and the asymptotic efficiencies of user j can vary with time. Therefore, to track these variations, we would initialize the TP algorithm using $a_m = a/m$. When a_m becomes small, we would use a small fixed value for a_m. For constructing the blind MMSE receiver, a similar technique is also used in [10].

6 Empirical results

We modeled a circular single cell of a dual-rate rate asynchronous DS/CDMA system. The radius of the cell was chosen to be $r_0 = 1000$ meters. It was assumed that the mobiles were uniformly distributed in the cell. This assumption yielded a probability density function $f(r) = 2r/r_0^2$ for the distance of a user from the base station. We used a path loss exponent $\alpha = 4$. The height of the base station was 30 meters so that the uplink channel gain of user j, h_j was $1/\left(r^2 + 30^2\right)^{\alpha/2}$. It was assumed that there were five low rate and five high rate users in the system. The processing gain of the low rate users was 64 and the processing gain of the high rate users was 8. The signature waveforms were generated using random signature sequences. The time offsets of different users were randomly generated. We assumed that the time offsets were multiples of the chip duration. The Gaussian noise had power spectral density 6×10^{-14} W/Hz.

In the experiment, user j was a high rate customer. The length of the observation window was $(n_j + d_j + 1)$, where n_j denoted the number of desired users' bits into the past and d_j is the number of its bits into the future and $n_j = d_j = 7$. After an interval of 8 bits, user j observed the same pattern of the signature waveforms from the other interfering users. Thus, the receiver dedicated to user j needed to know the specifications of 8 modified correlators and as a result, user j experienced 8 different asymptotic efficiencies. In our experiments, the following asymptotic efficiencies were used

$$\left\{\zeta_j^{(i)} | 0 \leq i \leq 7\right\} = \left\{0.3573 \;\; 0.3319 \;\; 0.3392 \;\; 0.3782 \;\; 0.4543 \;\; 0.3298 \;\; 0.3329 \;\; 0.2765\right\}$$

We assumed that the user j applied the AB link quality objective. The target average BER q_j^* was 10^{-2} and the corresponding target received power E_j^* was 9.5356×10^{-13}. The technique shown in Figure 1 was used to satisfy the requirement of iid noise vectors $\mathbf{N_j}(m)$.

In the first experiment, the average BER was determined as a function of the number of power control iterations for fixed a_m as well as $a_m = 10^{13}/m$; see Figure 2. Here in each iteration, the output of modified correlator 0 was used to update the transmitted power or $J = 1$. The mobile initialized the transmitted power from zero or $P_j(0) = 0$. The empirical results showed that for $a_m = 10^{13}/m$, the mobile achieved its QoS objective quickly compared to the fixed a_m of 0.1×10^{13}, or 0.01×10^{13}. These large values of a were needed, because the $|Y_j[P_j, \mathbf{N_j}(m)] - Y_j^*|$ of equation (15) were very small and $P_j^* \gg P_j(0)$. In this experiment, we found that in the case of fixed $a_m = a$, smaller values of a suffered from slow convergence, as suggested in the discussion of section 5. In the following experiments, we studied only the sequence $a_m = a/m$.

In Figure 3, we plot the average BER as a function of the number power control

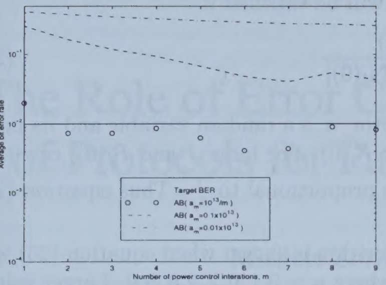


Figure 2: The average BER versus m for fixed a_m as well as $a_m = 10^{13}/m$. Here $J = 1$ and $P_j(0) = 0$.

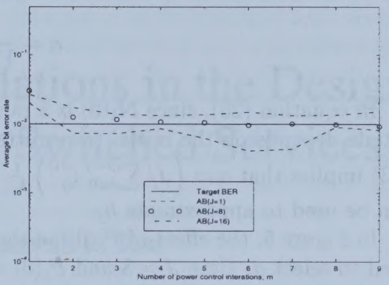


Figure 3: The average BER versus m for variable J. Here $a_m = 10^{13}/m$ and $P_j(0) = 0$.

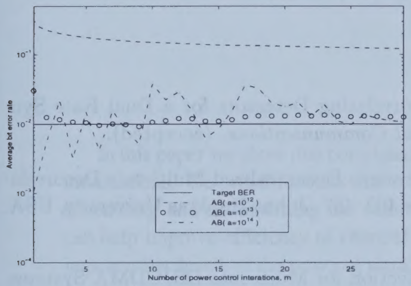


Figure 4: The average BER versus m for variable a when $a_m = a/m$. Here $P_j(0) = 0$.

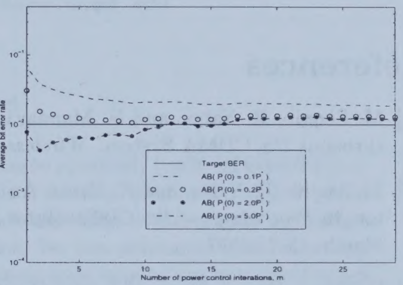


Figure 5: The average BER versus m for various $P_j(0)$, when equation (25) was used to select a. Here $J = 8$.

iterations for $J = 1, 8, 16$, where $a = 10^{13}$ and $P_j(0) = 0$. In this experiment, we observed that for $J = 8$, the TP algorithm converged quickly. However, for $J = 16$, the mobile increased the transmitted power uniformly to achieve q_j^*. Similar results were also observed in [11] where stochastic power control method for the matched filter receiver was studied.

Figure 4 describes the effect of a on the convergence when $J = 8$ and $P_j(0) = 0$. As expected, our empirical results showed that smaller values of a resulted in a slower convergence speed. On the other hand, larger values of a led to a wide variation in the initial transmitted power. These experimental results implied that neither a large a nor a small a was desirable for the mobile. This practical situation leads us to propose the following heuristic procedure to select the parameter a.

Let us assume that $P_j(0) > 0$. Using the target average power, Y_j^* and equation (13), we can estimate or approximate the target transmitted power, P_j^* as follows.

$$\hat{P}_j^* = \frac{Y_j^*}{Y_j[P_j(0), \mathrm{N_j}(0)]} P_j(0) \tag{23}$$

where $\hat{P}_j^*$ is the estimate of P_j^*. In order to be in the vicinity of P_j^* in the very first iteration, we will choose a so that

$$P_j(1) = \hat{P}_j^* \tag{24}$$

Equations (15) and (23) imply that $P_j(1) = \hat{P}_j^*$ will be satisfied if

$$a = \frac{P_j(0)}{Y_j[P_j(0), \mathrm{N_j}(0)]} \tag{25}$$

In equation (25), since $\mathrm{N_j}(0)$ is a random vector, a is a random variable and its magnitude depends on the initial transmitted power $P_j(0)$. For large J and $P_j(0)$, equation (13) implies that $a \approx \left(J / \sum_{i=0}^{J-1} \zeta_j^{(i)}\right) \frac{1}{h_j}$ which is proportional to $\frac{1}{h_j}$. Thus equation (25) can be used to approximate h_j.

In Figure 5, the effect of $P_j(0)$ on the TP algorithm is shown when equation (25) was used to select a. Here $J = 8$ and $P_j(0) = \kappa P_j^*$, where $\kappa = 0.1, 0.2, 2.0, 5.0$. Larger values of $P_j(0)$ yielded better estimates of P_j^* or smaller values of $|P_j(1) - P_j^*|$. As a result, the convergence speed of the power control was improved as $P_j(0)$ increased. For all values of $P_j(0)$, we observed that the mobile updated the transmitted power uniformly to achieve q_j^*.

References

[1] M. Saquib, R. Yates, and N. Mandayam. Decorrelating Detectors for a Dual Rate Synchronous DS/CDMA System. *Wireless Personal Communications.* (accepted).

[2] M. Saquib, R. Yates, and A. Ganti. An Asynchronous Decentralized Multi-rate Decorrelator. In *Proceedings of the CISS*, volume 1, pages 462–467, Johns Hopkins University, USA, March, 19-21 1997.

[3] U. Mitra. Observations on Jointly Optimal Detection for Multi-rate DS/CDMA Systems. In *Proceedings of 4th IEEE Communication Theory Mini-Conference, Global Telecommunications Conference*, UK, November 18 - 22 1996.

[4] M. L. Honig and S. Roy. Multi-user Communication with Multiple Symbol Rates. In *Proceedings of the IEEE Symposium on Information Theory, Whistler, BC, Canada*, page 381, September 1995.

[5] R. Lupas and S. Verdú. Near-Far Resistance of Multiuser Detectors in Asynchronous Channels. *IEEE Transactions On Communications*, 38(4):496–508, April 1990.

[6] S. Ulukus and R. Yates. Adaptive Power Control with MMSE Multiuser Detectors. In *Proceedings of ICC*, Montreal, Canada, 1997.

[7] M. Varanasi. Power Control for Multiuser Detection. In *Proceedings of the 30th Annual Conference on Information Sciences and Systems*, Princeton University, March 1996.

[8] D. J. Sakrison. Stochastic Approximation: A Recursive Method for Solving Regression Problems. *Advances in Communication Systems*, 2:51–106, 1966.

[9] M. Saquib, R. Yates, and A. Ganti. Quality of Service for an Asynchronous Multi-rate Decorrelator, Part II : Power Control. *IEEE Transactions on Communications.* (submitted).

[10] M. Honig, U. Madhow, and S. Verdú. Blind Adaptive Multiuser Detection. *IEEE Transactions on Information Theory*, 41(4):944–960, July 1995.

[11] S. Ulukus and R. D. Yates. Stochastic Power Control for Cellular Radio Systems. In *Proceedings of 34th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, September, 29 - October, 1 1996.

The Role of Error Correlations in the Design of Protocols for Packet Switched Services

Michele Zorzi and Ramesh R. Rao

Center for Wireless Communications, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0407
e-mail: {zorzi,rao}@ece.ucsd.edu

Abstract

In this paper we show that correlated errors can be potentially beneficial in packet communications. We present results that confirm this at the MAC level by focusing on packet scheduling and documenting the extent to which knowledge of the channel correlations can help improve efficiency of channel utilization. We then summarize some recent findings at other levels of the protocol stack, and discuss some implications to protocol design for the wireless channel.

1 Introduction

Fading impacts every layer of a protocol stack, regardless of the number and nature of functions assigned to the various layers. In this paper, we focus on the effect of error correlation introduced by fading, and in particular on its impact at the media access control (MAC) layer and the transport layer.

The traditional role of MAC protocols is to enable a group of uncoordinated users to share a common channel. In the development of MAC protocols for the wireless environment, one detects a certain convergence of thinking on the structure of such protocols. There is typically a segment of the protocol that deals with the initial signaling between the multiple users and the base station or central controller. Then there is a segment of the protocol that allocates the available resource, be it time slots, frequency channels or codes, to the users that have already signaled their needs. The allocation process is sometimes hard wired, as in the PRMA protocol [1] which specifies exactly how to allocate the slots. Other proposal, such as the DQRUMA protocol [2] or the CDPA architecture [3], are more flexible and allow for various types of allocation strategies to be implemented. In any event, nearly all recently proposed MAC protocols focus on the need to develop explicit or implicit scheduling algorithms for accommodating heterogeneous users but many of them ignore the capriciousness of the wireless channel that results from fading and mobility.

Some representative results on scheduling for the wireless channel are discussed next. Although some of these studies were made in the context of Link and Transport layer protocols, they involve ideas that are relevant across the various layers of the stack.

In the context of transport layer protocols, Badrinath and Sudame [4] examined the effect of mobility on UDP experimentally. The key idea was to add a certain modicum of flow control at the device driver level. Transmissions were suspended when the channel was estimated to be in a bad state. Two approaches to the estimation of the channel state were studied. In the "simple scheme" the estimates were based on measured signal strength of the previously received packet. This scheme, which largely ignores correlation patterns in the error process, was not found to be very satisfactory. In a more sophisticated "Hysteresis based" scheme, the authors based the estimate of the channel state on examining the received power levels of all packets received within a certain window. One particular "three out of five" scheme was found to be quite a bit more effective than the simpler scheme. Although the authors appear not to explicitly articulate the significance of error correlations, their results certainly reinforce the importance of error correlations. The idea of suspending transmissions when the channel is bad has also been explored in [5], where a probing technique is used for channel estimation.

Bhagwat and others [6] proposed a "Channel State Dependent Packet Scheduling" protocol that takes into account the nature of the wireless channel outages. Their simulations indicate that if perfect knowledge of the channel status is available, a 10 to 15 % increase in utilization accrues. Furthermore, Round Robin and Earliest Deadline first service policies were found to provide fairer service than FIFO, whereas the Longest Queue First policy was less fair. When no channel information is assumed to be known the authors implemented an implicit form of deferral by skipping over the "marked as bad" nodes and found that the Round Robin service policy continued to perform well. Once again we see the implicit role of tracking correlations in channel status.

Hara and others [7] have proposed a stop and wait protocol that updates packet lengths (within a set of geometrically related integers) based on the number of ACKs and NACKs received within a certain optimized observation period. They report high throughput under fast fading conditions.

In this paper, our overall goal is to study the extent to which knowledge of channel outage statistics can help improve efficiency. Towards this end, we present some results that describe the potential gain in throughput that can be achieved through the use of channel sensitive scheduling policies in MAC protocols. We also summarize some recent results that document the sensitivity of the system performance to correlation in channel errors at various layers, leading to some considerations about the protocol design criteria to be used in future generation wireless networks.

2 Traffic Model

We consider two classes of users, called *periodic* and *bursty*, respectively. Users in the same class are assumed to generate packets according to the same statistical patterns, whereas different classes have different patterns. All users in all classes generate packets which are equal in size and whose transmission takes one slot. We assume that N_p periodic users and N_b bursty users are present in the system.

The periodic users are either ON or OFF, according to a Markov process. This process can be characterized by the average ON and OFF times, T_{ON} and T_{OFF}. When in the ON state, a periodic user generates exactly one packet every Δ slots. When in the OFF state, no packets are generated. The channel load produced by one such user can be computed as $\rho_p = (T_{ON}/\Delta)/(T_{ON} + T_{OFF})$. This class of users can model voice connections with speech activity detection, as well as other VBR sources.

The bursty users generate one message in a slot with probability λ, and no messages with probability $1 - \lambda$ (Bernoulli arrival process). Every message is composed of a random number of packets. In our computations, the number of packets in a message is a geometrically distributed strictly positive random variable with mean μ. The channel load produced by a bursty user is $\rho_b = \lambda\mu$.

As both classes of users generate traffic in bursts, it makes sense to implement some reservation mechanism whereby, when a periodic user enters the ON state or a bursty user generates a message, a request is made to the central controller (base station, BS) over a signaling channel. When the BS becomes aware that a user has traffic to send, the BS includes it in the *scheduling set*, i.e., the set of users who need slots. For simplicity, we focus on the slot scheduling policy, and model the reservation scheme through a random delay between a message generation (or talkspurt initiation) and the time at which this is known at the BS.

Traffic is assumed to be delay-sensitive, so that packets have a finite lifetime, L. More specifically, when a packet is generated, it is stamped with its date of birth (DOB) and with its "date of death" DOD = DOB $+L$, after which the packet is discarded. In every slot, the queues are checked to see if any of the existing packets has reached its DOD, in which case it is dropped. Note that, in the case of bursty users, all packets in a message have the same DOB and DOD, and therefore are dropped together. We are interested in evaluating the packet (message) dropping probability, P_p (P_b), for a periodic (bursty) user.

3 Scheduling

In every slot, based on the available information, the BS makes a decision about which user will be assigned the channel for transmission. We assume that the signaling is perfect, so that in fact the designated user will transmit and all others will not.

As a first step towards studying more sophisticated scheduling algorithms, we assume that the decision about slot assignment is made instantaneously based on the status of the queues at all mobiles. More specifically, the BS looks at the queues of all users, and assigns some priority to each of them. The user with the highest priority will be assigned the slot and will transmit.

Even though collisions do not occur in our model, channel errors may corrupt the transmission. A failed packet will not leave the queue and will be retransmitted later. Here, we use a binary process to describe packet successes and failures due to fading and noise. Following [8, 9], we model this binary process for each user as a two-state Markov chain, which is identified by a pair of parameters, namely the average packet error rate, ε, and the average length of a burst of consecutive packet errors, b. Different users will experience different multipath and noise effects, and therefore the Markov processes which describe their channels will be independent of each other. For simplicity, we will assume that all users have the same statistical channel description, i.e., ε and b are the same for all users.

Under this general framework, there are many ways in which the BS can assign priorities to the users. In this preliminary study, we considered two, as described in the following. Besides being stamped with DOB and DOD, packets also contain a field which we call *scheduled date (SD)*, which is the time by which the system should try to deliver the packet[1]. This is needed for bursty users, where many packets have the same DOD, and therefore if the system tried to schedule them based on the DOD, then it would end up starting message transmission too late, so that message dropping will occur with high probability. To prevent this, packets in a message are assigned SD uniformly distributed between the time of arrival at the BS (i.e., the time at which the request is received) and the common DOD. This is of course not necessary in periodic traffic, where simultaneous arrivals are not allowed and no packets with the same DOD exist. In this case, DOD and SD are the same.

In slot t, the residual time to scheduled date, $SD - t$, is computed for the packet at the head of the queue[2], and the priority p is assigned according to the following rule:

$$p = \begin{cases} (SD - t)^{-1} & \text{if } SD - t > 0 \\ 2 & \text{otherwise} \end{cases} \tag{1}$$

Therefore, highest priority is given to the packets with the earliest SD. Note that the residual time to SD, $(SD - t)$, can be negative because in general $SD \leq DOD$, so that even if the scheduled date has passed the packet might not have been dropped yet. In this case, of course, high priority is given to that packet.

In the first scheduling scheme, which we will call *without channel information (CI)*, a slot is assigned to the user whose p is highest. If more than one users have the same p, the slot is randomly assigned to one of them.

In the second scheduling scheme we consider, called *with CI*, the BS is assumed to have perfect information about the status of the channel, and to give zero priority to all users whose channel is bad. Therefore, the previous scheme is applied only to those users who can successfully transmit, and no failed transmissions ever occur in this case. This scheme is obviously an idealization, but is useful in determining if using channel information can significantly enhance the performance. In environments where the channel is slowly varying and some prediction can be performed, it is certainly worth designing the protocols to take advantage of this information. Schemes based on various prediction strategies can be envisioned and will be the subject of future studies.

4 Numerical Results

In this section, we present some sample results for two cases, namely, only periodic users and mixed periodic/bursty traffic.

---

[1] Note that DOB, DOD and SD are not actual fields in the packets, but rather parameters which are known at the BS after a request is made. They are therefore fields in a virtual queue of slot requests which is kept at the BS for each active users, and do not constitute communications overhead in general.

[2] Queues are ordered by increasing DOD.

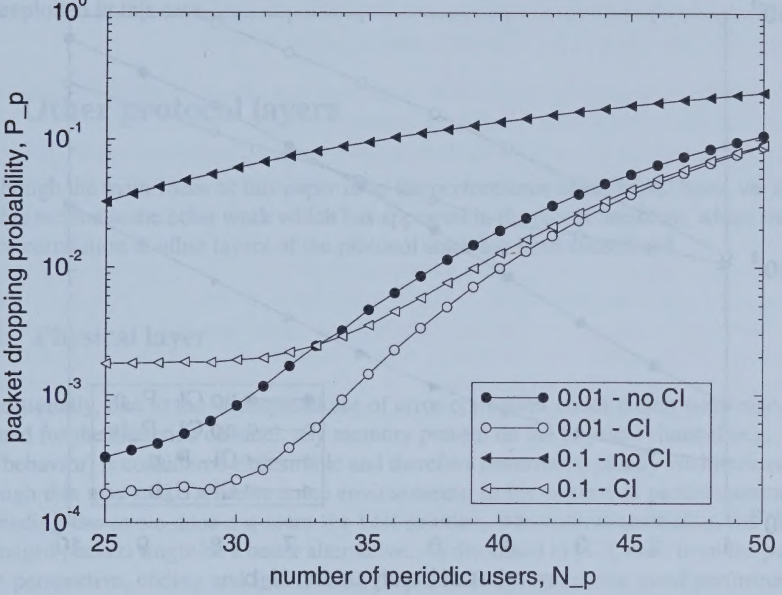


Figure 1: Packet dropping probability, P_p, vs. the number of admitted periodic users, N_p, for $\varepsilon = 0.1$ and 0.01, $b = 10$ slots, with and without channel information, no bursty users.

Figure 1 refers to the case in which only periodic users are present. We adopt the same model used in the study of Packet Reservation Multiple Access (PRMA) [10], where $T_{ON} = 1$ s, $T_{OFF} = 1.35$ s, the packet length is 576 bits (corresponding to 16 ms of speech), the channel data rate is 720 kbps so that one slot is 0.8 ms long and an active users generates a packet every 20 slots. The packet lifetime is assumed to be equal to twice the generation period, so that $L_p = 40$ slots. In this case, every user contributes a load of $\rho_p = 0.021$.

Results are plotted in terms of packet dropping probability, P_p, vs. the number of admitted users, N_p. Two values of the average packet error rate are shown, i.e., $\varepsilon = 0.1$ and 0.01, whereas the average length of a burst of errors is $b = 10$ slots. Two curves are given for each value of ε, for the two scheduling strategies with and without CI.

It can be seen that perfect channel information can improve performance significantly, especially for the higher value of the average error rate. Since the curves are flat, the gain in terms of number of admissible users (i.e., the horizontal distance between the curves) is large. Also, it should be noted that in both cases the scheme with CI can admit 40 users with a target dropping probability of 0.01 [10]. Even the scheme without CI, for $\varepsilon = 0.01$, can admit up to 37 users. Note that this outperforms PRMA, which can admit 37 users *in the absence of errors*, and does not work properly in the presence of errors. This shows that having a flexible scheduling instead of a simple reservation ALOHA helps, especially in an error-prone environment. A modified version of PRMA which allows fast retransmission has been proposed in [11], but the analysis there assumes independent errors, and this explains the extremely good results

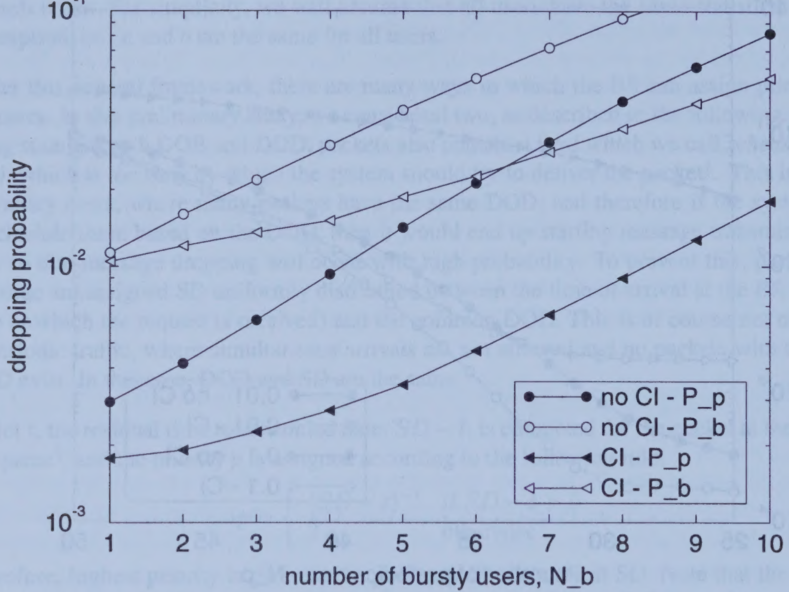


Figure 2: Packet and message dropping probabilities, P_p, P_b, vs. the number of admitted bursty users, N_b, for $N_p = 5$, $\varepsilon = 0.1$, $b = 10$ slots, with and without channel information.

reported even in the presence of large error rates.

Also, it is interesting to note that the curves show an "error floor," i.e., even decreasing the traffic does not help below a certain value. This is because, due to fading and noise, packets may be dropped even in the absence of congestion. This is more evident for high error rates and long bursts.

In the case shown in Figure 2, the two traffic classes (periodic and bursty) are assumed to coexist. The number of periodic users (whose parameters are as in the previous subsection) is kept constant and equal to $N_p = 5$, and the results are plotted vs. the number of bursty users, N_b. We assume $\lambda = 0.01$ and $\mu = 5$, and the message lifetime is $L_b = 40$ slots.

Four curves are plotted, since both P_p and P_b are considered with and without CI. Recall that P_p is the *packet* dropping probability for a periodic user, whereas P_b is the *message* dropping probability for a bursty user.

It can be seen that bursty users suffer more in this scenario, since P_b is significantly larger than P_p. It should also be recalled that dropping a message of a bursty user means dropping on average μ packets. Again, it can be seen that the performance difference between the scheme with and without CI can be significant. The error floor is present in this case as well.

These results, as well as others not presented here, show that information about the channel memory can be used to effectively schedule transmissions and to make the MAC protocol robust to channel impairments. Therefore, channel memory is to be seen as an opportunity to be exploited in this case.

5 Other protocol layers

Although the main focus of this paper is on the performance at the MAC layer, we summarize in this section some other work which has appeared in the recent literature, where the effect of error correlation at other layers of the protocol stack has been considered.

5.1 Physical layer

Traditionally, due to the widespread use of error-correction codes which were essentially designed for the Gaussian channel, any memory present on the physical channel (e.g., bursty error behavior) is considered undesirable and therefore removed, typically via interleaving. Even though this may make sense in some environments, in the context of packet communications spreading the errors does not seem the best solution, whereas retransmitting the (few) badly damaged packets might be a better alternative. As discussed in [12], even from the packet error rate perspective, coding and interleaving may not always guarantee good performance under typical bursty error behavior.

5.2 Data-link layer

It is well-known that for classic data-link protocols, such as Go-Back-N, clustering of packet errors is beneficial due to the way the protocol operates, in which a single error causes a number of packets to be discarded in a row [13, 14]. Therefore, if the packet error process which is presented to the data-link layer contains correlation, better throughput performance may be expected.

In addition to this benefit, channel correlation can be exploited for other purposes, since it constitutes useful information about the channel behavior. As an example of this philosophy, Zorzi and Rao [5] proposed a modified Go-Back-N protocol where channel correlation and prediction are used to avoid transmissions when the channel is in bad state, so as to avoid failures and the consequent waste of limited battery power. Observation of erroneous transmissions triggers a *probing mode* in which low-power transmissions are performed in order to track the evolution of the channel status, with regular transmissions being resumed once good channel quality is restored.

5.3 Transport layer

Because of its widespread use, a number of studies have considered the effect of running TCP over a wireless link. Being designed for wired connections, TCP fails in general to work well over wireless channels beacuse of the much larger error rate which results because of channel impairments and mobility (see [15] and the references therein). Analytical studies of the effect of independent errors have been presented in [16, 17]. However, most of these studies fail to capture a key feature of the wireless channel behavior, i.e., its memory.

Analytical studies of the performance of TCP in the presence of correlated errors have been presented in [18, 19], where it is shown that correlated errors have a beneficial effect. An interesting result which we recently obtained is that, in the presence of correlated errors, different versions of TCP (e.g., Tahoe and Reno) do not show significant difference in performance. This is due to the fact that newer versions of TCP have been proposed based on the implicit assumption of a wired connection. As it turns out, in the presence of the error behavior typical of the wireless channel, their performance improvement vanishes [20], as was also observed in [21] for a periodically available channel. On the other hand, as shown in [19], an increased degree of correlation in the packet error process may correspond to a very significant performance enhancement (as large as a factor of 2 or 3 in throughput). This facts suggest that a clever design of the lower layers where memory in the packet error process is preserved may turn out to be much more effective than the development of more sophisticated TCP releases.

6 Implications for protocol design

A general conclusion which can be drawn from the results discussed in the above is that error burstiness may be desirable (to a certain extent), since it results in better performance of protocols such as data-link schemes or TCP, and since it provides additional information on the channel behavior which can be exploited. Classic techniques aimed at breaking the channel memory (e.g., interleaving), although justified by physical layer consideration (e.g., improved performance of error-correction codes), may turn out to be undesirable when a broader view is taken.

More generally, our results show that for a given average packet error rate, independent packet errors do not constitute the best error process, in the sense that some degree of correlation between consecutive errors may turn out to be beneficial. In this sense, being able to design the lower layers of a system in such a way that this burstiness (which is already present because of the nature of the wireless channel!) is somewhat preserved, is certainly desirable from the higher layers' point of view.

The main point we can make here is therefore twofold. On the one hand, a physical layer which passes up correlated packet errors is more desirable than one which decorrelates the channel artifacts. On the other hand, once the presence of this degree of correlation is granted at the packet level, there is room for cleverness in the protocol design, to achieve various objectives, e.g.: energy saving by avoiding transmission when the channel quality is degraded; optimal choice of the protocol parameters based on the second-order error statistics information; channel prediction for improved MAC performance by granting permission to transmit to users whose channel is best; no need for more elaborate or new TCP releases for connections comprising a wireless link.

Recognizing the usefulness of a better understanding of packet error statistics and of its value in making appropriate decisions at the higher layers may be a driver for improved "channel-aware" protocol design and for physical layer choices taking into account the needs of the whole protocol stack.

7 Conclusions

In this paper we have attempted to single out ideas and results that shed some light on the consequences of fading and especially the error correlation it introduces. It is clear that protocols and algorithms can be designed at almost every layer of a protocol stack to combat fading. One common theme that has emerged is the sensitivity of the performance at every layer to the higher order statistics of the error process. The understanding of the consequences of fading, as it manifests itself at the various layers, ought to foster more cross layer solutions to the challenges that mobility and fading present in the design of wireless networks.

References

[1] D.J. Goodman *et al.*, "Packet reservation multiple access for local wireless communications," *IEEE Trans. Commun.*, vol. 37, pp. 885–890, Aug. 1989.

[2] M.J. Karol, Zhao Liu; K.Y. Eng, "Distributed-queueing request update multiple access (DQRUMA) for wireless packet (ATM) networks," in *Proc. ICC'95*, pp. 1224–31, Jun. 1995.

[3] F. Borgonovo, M. Zorzi, L. Fratta, V. Trecordi, G. Bianchi, "The Capture-Division Packet Access (CDPA) for Wireless Personal Communications," *IEEE J. Selected Areas Comm.*, vol. SAC-14, pp. 609–623, May 1996.

[4] B. R. Badrinath, P. Sudame, "To Send or Not to Send: Implementing deferred Transmissions in a Mobile Host" in *Proc. IEEE 16th International Conference on Distributed Computing Systems*, pp. 327–333, 27–30 May 1996.

[5] M. Zorzi, R.R. Rao, "Error Control and Energy Consumption in Communications for Nomadic Computing," *IEEE Trans. Computers* (special issue on Mobile Computing), vol. 46, pp. 279–289, Mar. 1997.

[6] P. Bhagwat, P. Bhattacharya, A. Krishna, and K. Tripathi, "Using channel state dependent packet scheduling to improve TCP throughput over wireless LANs," *Wireless Networks,* pp. 91–102, 1997.

[7] S. Hara, A. Ogino, M. Araki, M. Okada and N. Morinaga, "Throughput Performance of SAW-ARQ Protocol with Adaptive Packet length in Mobile Data Transmission," in *IEEE Trans. Veh. Tech.* pp. 561–569, Aug. 1996.

[8] M. Zorzi, R. R. Rao, and L. B. Milstein, "On the accuracy of a first-order Markov Model for data transmission on fading channels," *Proc. IEEE ICUPC'95,* pp. 211–215, Nov. 1995.

[9] M. Zorzi, R.R. Rao, L.B. Milstein, "A Markov model for block errors on fading channels," in *Proc. PIMRC'96*, pp. 1074–78, Oct. 1996.

[10] S. Nanda, D. J. Goodman, and U. Timor, "Performance of PRMA: A Packet Voice Protocol for Cellular Systems", *IEEE Trans. Veh. Tech.*, vol. 40, pp. 584–598, Aug. 1991.

[11] G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci, M. Zorzi, "C-PRMA: a Centralized Packet Reservation Multiple Access for Local Wireless Communications," *IEEE Trans. Veh. Tech.*, vol. 46, pp. 422–436, May 1997.

[12] M. Zorzi, R.R. Rao, "On the Statistics of Block Errors in Bursty Channels," *IEEE Trans. Comm.*, vol. 45, pp. 660–667, Jun. 1997.

[13] C.H.C. Leung, Y. Kikumoto, S.A. Sorensen, "The throughput efficiency of the Go-Back-N ARQ scheme under Markov and related error structures", *IEEE Trans. Commun.*, vol. COM-36, pp. 231–234, Feb. 1988.

[14] M. Zorzi, R.R. Rao, "Throughput analysis of ARQ Go-Back-N protocol in Markov channels with unreliable feedback," in *Proc. IEEE ICC'95*, pp. 1232–37, Jun. 1995.

[15] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improvong TCP performance over wireless links," *Proc. ACM SIGCOMM'96*, Stanford, Aug. 1996.

[16] A. Kumar, "Comparative performance analysis of versions of TCP in a local network with a lossy link," *Tech. Rep. WINLAB-TR-129*, Oct. 1996.

[17] T. V. Lakshman, U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random loss," *IEEE/ACM Trans. on Networking*, pp. 336–350, Jun. 1997.

[18] A. Kumar, and J. Holtzman, "Comparative performance analysis of versions of TCP in a local network with a lossy link, Part II: Rayleigh fading mobile radio link," *Tech. Rep. WINLAB-TR-133*, November 1996.

[19] M. Zorzi, R. R. Rao, "Effect of correlated errors on TCP," in *Proc. 1997 CISS*, Baltimore, MD, March 1997.

[20] A. Chockalingam, M. Zorzi, R.R. Rao, "Performance of TCP Reno on wireless fading links with memory," submitted to *IEEE ICC'98*, Aug. 1997.

[21] Mooi Choo Chuah, On-Ching Yue, Antonio DeSimone, "Performance of two TCP implementations in Mobile Computing Environments" in *Proc. IEEE GLOBECOM'95*, pp. 339–344, Nov. 1995.

TCP Improvements for Heterogeneous Networks: The Daedalus Approach

Hari Balakrishnan*, Venkata N. Padmanabhan*, Srinivasan Seshan+, Mark Stemm*, Elan Amir*, Randy H. Katz*
`{hari,padmanab,stemm,elan,randy}@cs.berkeley.edu, srini@watson.ibm.com`

**Computer Science Division*
University of California at Berkeley
Berkeley, CA 94720

+IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

Abstract

In this paper, we present a summary of the research work done by the Daedalus group at the University of California at Berkeley to improve TCP performance in a variety of situations, including lossy wireless networks, asymmetric networks, and Web traffic workload. To improve the performance of TCP over cellular wireless networks, we have implemented a TCP-aware link layer protocol called the *snoop protocol* that isolates wired senders from the lossy characteristics of a wireless link. To improve TCP performance over networks exhibiting asymmetry, we have designed and evaluated techniques that reduce the adverse effects of the constrained reverse path, which TCP acks traverse. To improve TCP performance for Web-like traffic, we have designed and evaluated an enhanced data-driven loss recovery scheme that improves the performance of short transfers, as well as an *integrated congestion control and loss recovery* approach that improves the end-to-end performance of multiple, short TCP connections from a single host.

1. Introduction

The growth of the Internet and the World-Wide Web has increased the diversity of situations in which TCP connections are used. There are two aspects to this diversity: heterogeneous networks and new traffic workload. Networks with wireless links and asymmetric networks such as cable modem, direct broadcast satellite and packet radio networks are increasingly being used to provide Internet access to the workplace and home. Also, the preponderance of Web traffic has resulted in most TCP connections being relatively short in duration. These trends have impacted TCP significantly.

First, the traditional assumption made by TCP that packet losses are indicators of network congestion is correct for wired networks but breaks down in wireless networks where packets are often corrupted by bit errors. Second, the design of TCP algorithms heretofore has primarily focussed on the network path characteristics on the direction of data transfer. The flow of TCP acknowledgments has received only limited attention (e.g., [23]). However, in the presence of bandwidth and other types of asymmetry, the constrained flow of acks can have a significant adverse effect on performance. Finally, TCP's window growth and data-driven loss recovery algorithms have been designed with long data transfers in mind. Their effectiveness is often limited in the context of short Web transfers. In summary, TCP falls short on several fronts in the presence of new types of networks and traffic workloads.

The Daedalus group at the University of California at Berkeley [11] has been working on improving TCP performance in each of the contexts mentioned above: wireless links, asymmetric networks and Web traffic. This paper presents a brief summary of the techniques that we have investigated and our findings:

- To improve performance over wireless links, we have implemented a TCP-aware link layer protocol called the *snoop protocol* that shields a TCP sender from the lossy characteristics of the wireless link [6, 7]. We have also compared this approach to a variety of other approaches for improving wireless TCP performance[4].
- To improve performance over asymmetric links, we have implemented techniques for eliminating the acknowledgment (ack) bottleneck that comes about as a result of *bandwidth asymmetry*, when the forward link has significantly higher bandwidth than the reverse link [3].
- To improve the performance of TCP connections for Web traffic, we have proposed an enhanced loss recovery scheme that improves performance of short transfers and an integrated connection approach to loss recovery and congestion control that improves performance for concurrent Web connections [5].

The rest of this paper is organized as follows. In Section 2 we describe our improvements for lossy wireless links. In Section 3 we compare this approach to other approaches for improving TCP performance over wireless links. In Section 4 we describe our improvements for overcoming asymmetries in network characteristics between the forward and reverse directions. In Section 5 we describe our improvements for Web traffic and conclude in Section 6.

2. The Snoop Protocol [6]

Reliable transport protocols like TCP [10, 20, 8] have been tuned for traditional networks made up of wired links and stationary hosts. Due to the relatively low bit-error rates over wired networks, all packet losses are assumed to be due to congestion. In the presence of the high error rates and periods of intermittent connectivity characteristic of wireless links, TCP reacts to packet losses as it would in the wired environment: it drops its transmission window size before retransmitting packets, initiates congestion control or avoidance mechanisms (e.g., slow start [14]) and resets the retransmission timer (Karn's Algorithm [16]). These measures result in a unnecessary reduction in the link's bandwidth utilization, thereby causing a significant degradation in performance in the form of poor throughput and very high interactive delays[9].

In [7], we present the details of the *snoop* protocol, a protocol designed to improve performance for data transfers to and from the mobile host. Here, we outline the basic schemes used to improve performance for data transfers to the mobile host.

2.1 Details of the Snoop Protocol

The snoop protocol is a TCP-aware link protocol, running primarily at the base station to the wireless network. The snoop agent at the base station monitors every TCP packet that passes through the

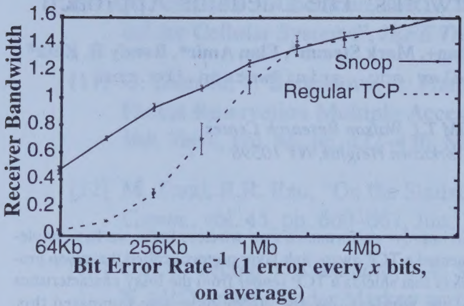


Figure 1. Throughput received by the mobile host at different bit-error rates ($\log_2$ scale).

connection in either direction. The snoop agent maintains a cache of TCP packets sent from the Fixed Host (FH) that haven't yet been acknowledged by the Mobile Host (MH). This is possible because TCP has a cumulative acknowledgment policy for received packets. When a new packet arrives from the FH, the snoop agent adds it to its cache and passes the packet on to the routing code which performs the normal routing functions. The agent also keeps track of all the acknowledgments sent from the mobile host. It detects the loss of packets either by the arrival of duplicate acknowledgments, or by a local timeout. Upon detecting packet loss, it retransmits the lost packet to the mobile host if it has the packet cached. In addition, the basestation hides the packet loss from the sender by not propagating duplicate acknowledgments, thereby preventing unnecessary congestion control mechanism invocations at the source, which would other wise result because of TCP's fast retransmission mechanism.

2.2 Results

We performed several experiments with the snoop protocol on our wireless testbed, consisting of Pentium PC base stations and IBM ThinkPad mobile hosts communicating over a 915 MHz AT&T Wavelan [21], all running BSD/OS 2.0, and compared the resulting performance with unmodified TCP. In the presence of no packet losses, the wireless link was able to support about 1.5 Mbits/s out of a rated maximum of 2 Mbits/s.

A preliminary design of the snoop protocol with some simulation results appeared in [1]. To measure the performance of the implementation under controlled conditions, we used a Poisson-distributed bit error model. Figure 1 compares the bandwidth of a connection using the snoop protocol with that of a connection using an unmodified TCP implementation for various Poisson-distributed bit-error rates shown on a log scale. We generated a Poisson distribution for each bit-error rate and changed the TCP checksum of the packet at the basestation if the error generator determined that the packet should be dropped at the receiver before forwarding the packet over the wireless link.

We see that for error rates of over 10^{-6} (close to the 1 Mb point on the x-axis of the graph) the snoop protocol performs significantly better than unmodified TCP, achieving a bandwidth improvement factor of 2 to 20 depending on the bit error rate. In fact, the snoop protocol was robust and completed the run even when every other packet was being dropped over the last link, while the regular TCP connection didn't make any progress. Under conditions of very low bit error rates ($< 10^{-6}$), there does not seem to be any significant difference between the snoop and regular TCP implementations. At such low bit errors the snoop layer is essentially just a store-and-forward agent, and incurs no additional overhead.

In the next section, we compare snoop to other link-layer approaches for improving TCP performance on high-loss wireless networks in order to understand precisely why the protocol performs well.

3. Comparison of Snoop with Other Approaches

Several schemes have been proposed to the alleviate the effects of non-congestion-related losses on TCP performance over networks that have wireless or similar high-loss links [2, 7, 22]. These schemes choose from a variety of mechanisms, such as local retransmissions, split-TCP connections, and forward error correction, to improve end-to-end throughput. However, it is unclear to what extent each of the mechanisms contributes to the improvement in performance. In [4], we examine and compare the effectiveness of these schemes and their variants, and experimentally analyze the individual mechanisms and the degree of performance improvement due to each.

We break the approaches into three categories: (1) *link-layer protocols*, ones that operate solely at the link layer, (2) *end-to-end protocols*, ones that operate solely at the connection endpoints, and (3) *split-connection protocols*, ones that break the TCP connection at the wireless endpoints.

We seek to answer the following specific questions:

1. What combination of mechanisms results in the best performance for each of the protocol categories?
2. How important is it for link-layer schemes to be aware of TCP algorithms to achieve high end-to-end throughput?
3. How useful are selective acknowledgments in dealing with lossy links, especially in the presence of burst losses?
4. Is it important for the end-to-end connection to be split in order to effectively shield the sender from wireless losses and obtain the best performance?

In the next section, we focus on the importance of transport-awareness in link-layer protocols.

3.1 Results

The results reported in this section were obtained through implementing and evaluating the various protocols in a wireless testbed consisting of Pentium PC base stations and IBM ThinkPad mobile hosts communicating over a 915 MHz AT&T Wavelan [21], all running BSD/OS 2.0.

In order to measure the performance of the protocols under controlled conditions, we generate errors on the lossy link using a Poisson-distributed bit-error model. The receiving entity on the lossy link generates a Poisson distribution for each bit-error rate and changes the TCP checksum of the packet if the error generator determines that the packet should be dropped. Losses are generated in both directions of the wireless channel, so TCP acknowledgments are dropped too, albeit at a lower per-packet rate. For most of the experiments, the TCP data packet size is 1400 bytes and the average error rate is one every 64 KBytes (this corresponds to a bit-error rate of about $1.9\mathrm{x}10^{-6}$). The choice of the Poisson-distributed error model is motivated by our desire to understand the precise dynamics of each protocol in response to a wireless loss, and is not an attempt to empirically model a wireless channel. While the actual performance numbers will be a strong function of the exact error model, the relative performance is dependent on how the protocol behaves after one or more losses in a

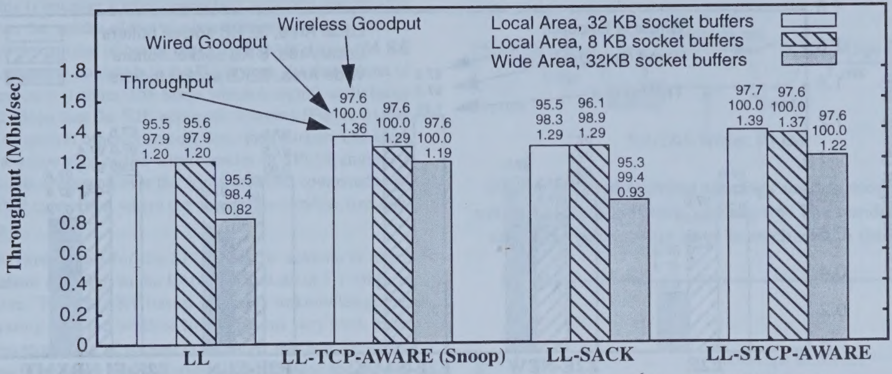


Figure 2. Performance of link-layer protocols: bit-error rate = 1.9×10^{-6} (1 error/65536 bytes).

single TCP window. Thus, we expect our overall conclusions to be applicable under other patterns of wireless loss as well.

In our experiments, we attempt to ensure that losses are only due to wireless errors (and not congestion). This allows us to focus on the effectiveness of the mechanisms in handling such losses. The WAN experiments are performed across 16 Internet hops with minimal congestion[1] in order to study the impact of large delay-bandwidth products.

Each run in the experiment consists of an 8 MByte transfer from the source to receiver across the wired net and the WaveLAN link. A particular metric of interest is the *goodput* of a path (or link) which is defined as the ratio of the actual transfer size to the total number of bytes transmitted over that path. In general, the wired and wireless goodputs differ because of wireless losses, local retransmissions and congestion losses in the wired network. For each protocol, we measured the end-to-end throughput and goodputs for the wired and one-hop wireless paths. The rest of this section presents and discusses the results of these experiments.

3.1.1 Link-Layer Protocols

The performance of the various link-layer protocols is summarized in Figure 2. In this section, we compare four protocols: (i) LL, a basic link layer protocol with cumulative acks and timer-based retransmissions, (ii) LL-TCP-AWARE, a link layer protocol that is aware of TCP (such as snoop), (iii) LL-SACK, a link layer protocol with selective acknowledgments, and (iv) LL-STCP-AWARE, a link layer protocol that combines TCP awareness with selective acknowledgments.

Traditional link-layer protocols operate independently of the higher-layer protocol, and consequently, do not necessarily shield the sender from the lossy link. This could adversely impact TCP performance for two reasons: (i) competing retransmissions caused by an incompatible setting of timers at the two layers, and (ii) the effect of the link layer protocol on the TCP fast retransmission mechanism. However, (i) is not the dominating effect when link layer schemes such as LL are used with TCP Reno and its variants. These TCP implementations have coarse retransmission timeout granularities that are typically multiples of 500 ms, while link-layer protocols typically have much finer timeout granularities. The real problem is that when packets are lost, link-layer protocols that do not attempt in-order delivery across the link (e.g., LL) cause packets to reach the TCP receiver out-of-order. This leads to the generation of duplicate acknowledgments by the TCP receiver, which causes the sender to invoke fast retransmission and recovery and can potentially cause degraded throughput and goodput, especially when the delay-bandwidth product is large.

Our results substantiate this claim, as can be seen by comparing the LL and LL-TCP-AWARE results. For a packet size of 1400 bytes and a bit error rate of 1.9×10^{-6} (1/65536 bytes), the packet error rate is about 2.3%. Therefore, an optimal link-layer protocol that recovers from errors locally and does not compete with TCP retransmissions should have a wireless goodput of 97.7% and a wired goodput of 100% in the absence of congestion. In the LAN experiments, the throughput difference between LL and LL-TCP-AWARE is about 10%. However, the LL wireless goodput is only 95.5%, significantly less than LL-TCP-AWARE's wireless goodput of 97.6%, which is close to the maximum achievable goodput. When a loss occurs, the LL protocol performs a local retransmission relatively quickly. However, enough packets are typically in transit to create more than 3 duplicate acknowledgments. These duplicates eventually propagate to the sender and trigger a fast retransmission and the associated congestion control mechanisms. These fast-retransmissions result in reduced goodput; about 90% of the lost packets are retransmitted by both the source (due to fast retransmissions) and the base station.

The effects of this interaction are much more pronounced in the wide area experiments — the throughput difference is about 30% in this case. The cause for the more pronounced deterioration in performance is the higher bandwidth-delay product of the wide-area connection. The LL scheme causes the sender to invoke congestion control procedures often due to duplicate acknowledgments and causes the average window size of the transmitter to be lower than for LL-TCP-AWARE.

In summary, our results indicate that a simple link-layer retransmission scheme could adversely impact TCP performance. An enhanced link-layer scheme, that uses knowledge of TCP semantics to prevent duplicate acknowledgments caused by wireless losses from reaching the sender, achieves significantly better performance.

3.1.2 End-To-End Protocols

In this section, we compare 5 protocols: (i) E2E, basic TCP Reno, (ii) E2E-NEW, TCP Reno enhanced with partial fast recovery, (iii) E2E-SACK, TCP Reno enhanced with selective acknowledgments, (iv) E2E-ELN, TCP Reno enhanced with Explicit Loss Notification

---

[1] WAN experiments were performed between 10 pm and 4 am, PST.

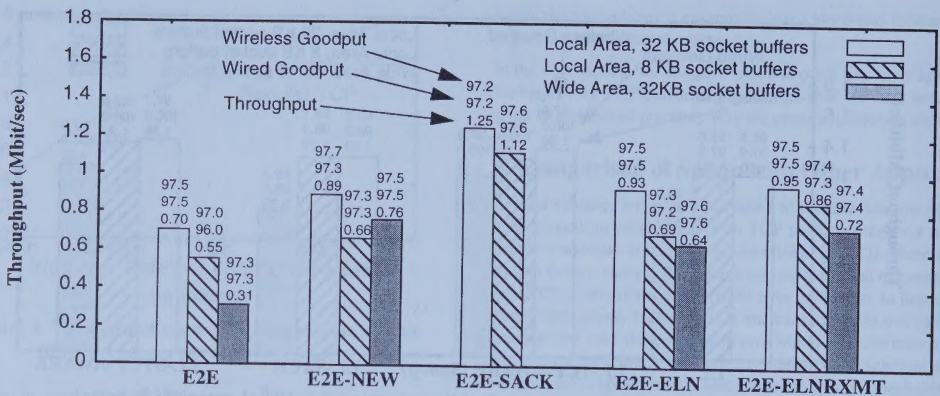


Figure 3. Performance of end-to-end protocols: bit error rate = 1.9×10^{-6} (1 error/65536 bytes).

(ELN), and (v) E2E-ELNRXMIT, an enhancement to E2E-ELN that immediately retransmits missing packets when receiving acknowledgments with the ELN bit set.

The performance of the various end-to-end protocols is summarized in Figure 3. The performance of unmodified TCP Reno, the baseline E2E protocol, highlights the problems with TCP over lossy links. With a 2.3% packet loss rate (as explained in Section 3.1.1), the E2E protocol utilizes less than 50% of the available bandwidth in the local area and less than 25% of the available bandwidth in the wide area experiments. However, all the end-to-end protocols achieve goodputs close to the optimal value of 97.7%. The primary cause for the low bandwidth is the large number of timeout-based retransmissions that occur during the transfer and the small average window size during the transfer that prevents the "data pipe" from being kept full and reduces the effectiveness of the fast retransmission mechanism.

The modified end-to-end protocols improve throughput by retransmitting packets known to have been lost on the wireless hop earlier than they would have been by the baseline E2E protocol, and by reducing the fluctuations in window size. The E2E-NEWRENO, E2E-ELN and E2E-SACK each use new TCP options and more sophisticated acknowledgment processing techniques to improve the speed and accuracy of identifying and retransmitting lost packets. E2E-NEWRENO, which uses partial acknowledgment information to recover from multiple losses in a window at the rate of one packet per round-trip time, performs between 10 and 25% better than E2E over a LAN and about 2.3 times better than E2E in the WAN experiments.

One way of eliminating long delays caused by coarse timeouts is to maintain as large a window size as possible. E2E-NEWRENO remains in fast recovery if the new acknowledgment is only partial, but reduces the window size to half its original value upon the arrival of the first new acknowledgment. The E2E-ELN and E2E-ELN-RXMT protocols use ELN (Explicit Loss Notification) information to prevent the sender from reducing the size of the congestion window in response to a wireless loss. Both these schemes perform better than E2E-NEWRENO, and over two times better than E2E. This is because there is usually enough data in the pipe to trigger a fast retransmission for E2E-ELN. The performance benefits of E2E-ELN-RXMT are more pronounced when the socket buffer size is smaller, as the numbers for the 8 KB socket buffer size indicate.

Finally, we also experimented with a simple SACK scheme based on a subset of the SMART proposal [16] in the local area. This protocol was the best of the end-to-end protocols in this situation, achieving a throughput of 1.25 Mbps (in contrast, the best local scheme, LL-OPT, obtained a throughput of 1.40 Mbps).

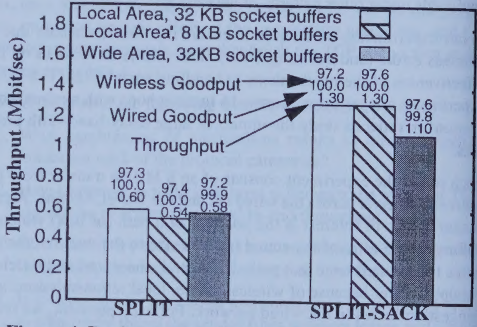


Figure 4. Performance of split-connection protocols: bit error rate = 1.9×10^{-6} (1 error/65536 bytes).

In summary, E2E-NEWRENO is better than E2E, especially for large socket buffer sizes. Adding ELN to TCP improves throughput significantly by successfully preventing unnecessary fluctuations in the transmission window. Finally, SACKs provide significant improvement over TCP Reno, but perform about 10-15% worse than the best local schemes in the LAN tests.

3.2 Split-Connection Protocols

In this section, we examine two algorithms: (i) SPLIT, where the connection is split at the base station into two TCP Reno connections, and (ii) SPLIT-SACK, where the connection is also split at the base station into two TCP connections, but the wired connection uses TCP Reno and the wireless connection uses TCP Reno enhanced with selective acknowledgments. The main advantage of the split-connection approaches is that they isolate the TCP source from wireless losses. The TCP sender of the second, wireless connection performs all the retransmissions in response to wireless losses.

Figure 4 shows the throughput and goodput for the split connection approach in the LAN and WAN environments. We see that the throughput achieved by the SPLIT approach (0.6 Mbps) is quite low, about the same as that for end-to-end TCP Reno (labeled E2E in

Figure 3). This is because a split-connection approach (eventually) stalls whenever the sender of the wireless connection experiences a timeout, since the amount of buffer space at the base station (64 KB in our experiments) is bounded. In the WAN case, the throughput of the SPLIT approach is about 0.58 Mbps which is significantly better than the 0.31 Mbps that the E2E approach achieves (Figure 3), but not as good as several other protocols described earlier. The large congestion window size of the wired sender in SPLIT enables a higher bandwidth utilization over the wired network, compared to an end-to-end TCP connection where the congestion window size fluctuates rapidly.

As expected, throughput for the SPLIT-SACK scheme is much higher. It is about 1.3 Mbps in the LAN case and about 1.1 Mbps in the WAN case. The SMART-based selective acknowledgment scheme operating over the wireless link performs very well, especially since no reordering of packets occurs over this hop. However, there are a few times when both the original transmission and the first retransmission of a packet get lost, which sometimes results in coarse timeouts. This explains the difference in throughput between the SPLIT-SACK scheme and the LL-OPT scheme (Figure 2).

In summary, while the split-connection approach results in good throughput if the wireless connection uses some special mechanisms, the performance does not exceed that of a well-tuned, TCP-aware link-layer protocol (LL-OPT). Moreover, the link-layer protocol preserves the end-to-end semantics of TCP acknowledgments, unlike the split-connection approach. This demonstrates that the end-to-end connection need not be split at the base station in order to achieve good performance.

In the next section, we focus on improving TCP performance when the wireless link exhibits asymmetric properties.

4. TCP over Asymmetric Networks

In this section, we focus on the challenges to end-to-end TCP performance that arise due to network asymmetry, especially in the context of wide-area wireless networks. The increased interest in asymmetric networks is motivated by technological and economic considerations as well as by popular applications such as Web access, where substantially more data flows towards the client (the *forward* direction) than from it (the *reverse* direction).

Examples of networks that exhibit asymmetry include *wireless cable modem* networks, *Direct Broadcast Satellite* (DBS) networks, and *Asymmetric Digital Subscriber Loop* (ADSL) networks, where bandwidths in the forward direction are often orders of magnitude higher than in the reverse. Such asymmetry is accentuated when the channel is uni-directional, necessitating the use of a different, often low-bandwidth channel (e.g., a dialup line or a wireless channel) for communication in the reverse direction.

As a general definition, we say that a network exhibits asymmetry with respect to TCP performance if the throughput achieved in one direction depends significantly on the link and traffic characteristics in the other direction. We focus on *bandwidth asymmetry* here, where the bandwidth in the forward direction is greater than the bandwidth in the reverse direction. We also briefly discuss asymmetry in *latency* and *media access*. A more detailed treatment appears in [3].

The simulation topology we used to investigate the effects of bandwidth asymmetry is shown in Figure 5. In the following sub-sections, we discuss several performance problems that we observed based on experiments conducted in a real testbed as well as in a simulator. We then discuss some solution techniques and evaluate their efficacy via simulations.

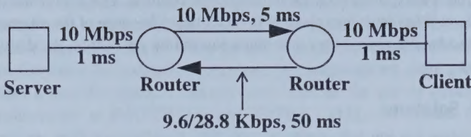


Figure 5. The simulation topology used to model a network with bandwidth asymmetry. The bandwidth and delay parameters have been chosen to closely

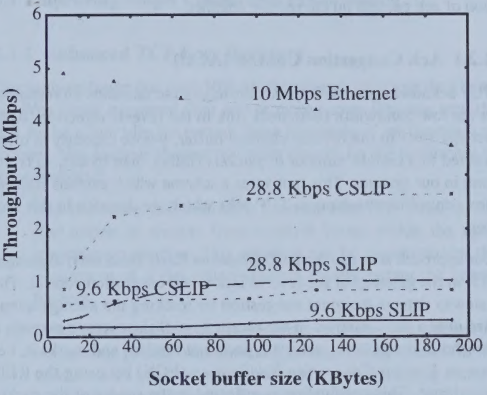


Figure 6. Performance of the wireless cable system using different return channels, across a range of socket buffer sizes.

4.1 Problems due to Bandwidth Asymmetry

The main problem with bandwidth asymmetry for unidirectional transfers is that the constrained reverse bandwidth can prevent TCP from achieving the full bandwidth of the forward path because of the slow arrival of acks. This is because TCP ack clocking, required for the connection to make steady progress, can break down. For example, consider two data packets transmitted by the sender in quick succession. While in transit to the receiver, these packets get spaced apart according to the bottleneck link bandwidth in the forward direction. The principle of ack clocking is that the acks generated in response to these packets retain the same spacing (in time) all the way back to the sender, enabling it to clock out new data packets with the same spacing.

However, the limited reverse bandwidth and consequent queuing effects could alter the inter-ack spacing. When acks arrive at the bottleneck link in the reverse direction at a faster rate than the link can support, they get queued behind one another. The spacing between them when they emerge from the link is dilated with respect to their original spacing. Thus the sender clocks out new data at a slower rate than if there had been no queuing of acks. Further, a point is reached when the reverse channel buffer gets filled up with acks. From that point on, acks often get dropped. As a consequence, the sender becomes bursty and its window growth is slowed down.

Figure 6 shows TCP throughput measurements from the Hybrid wireless cable modem network with various reverse channels. The x-axis shows the socket buffer size and the y-axis shows the throughput in Mbps. We see that ack dilation limits the throughput for dialup reverse channels without SLIP header compression. Header compression allows more acknowledgments to traverse the bottleneck link in the backward direction, increasing performance. The performance with an Ethernet return channel is much better because of the absence of bandwidth asymmetry and much lower link delay than the dialup lines.

4.2 Solutions

The problems discussed above arise because of contention for the bottleneck resources in the reverse direction — link bandwidth and buffer space. This observation serves as the starting point for the solutions discussed below. We present two techniques — ack congestion control and ack filtering — for alleviating the effects of congestion of ack packets on the reverse channel.

4.2.1 Ack Congestion Control (ACC)

TCP acknowledgments impose non-negligible demands on resources at the low-bandwidth bottleneck link in the reverse direction. Acks occupy slots in the reverse channel buffer, whose capacity is often limited to a certain number of *packets* (rather than bytes), as is the case in our systems. This motivates a scheme which extends congestion control mechanisms to TCP acks which we describe in this section.

Our approach is to use the RED (*Random Early Detection*) algorithm [13] at the gateway of the reverse link to aid congestion control. The gateway detects incipient congestion by tracking the average queue size over a time window in the recent past. If the average exceeds a threshold, the gateway selects a packet at random and marks it, i.e. sets an *Explicit Congestion Notification* (ECN) bit using the RED algorithm[2]. This notification is reflected to the sender of the packet by the receiver. Upon receiving a packet with ECN set, the sender reduces its sending rate.

The important point to note is that with ACC, *both* data packets and TCP acks are candidates for being marked. The TCP receiver maintains a dynamically varying delayed-ack factor, *d*, and sends one ack for every *d* data packets. When it receives a packet with the ECN bit set, it increases *d* multiplicatively, thereby decreasing the frequency of acks also multiplicatively. Then for each subsequent round-trip time (determined using the TCP timestamp option) during which it does not receive an ECN, it linearly decreases the factor *d*, thereby increasing the frequency of acks. Thus, the receiver mimics the standard congestion control behavior of TCP senders in the manner in which it sends acks.

There are bounds on the delayed-ack factor *d*. Obviously, the minimum value of *d* is 1, since at most one ack is sent per data packet. The maximum value of *d* is determined by the sender's window size, which is conveyed to the receiver in a TCP option. The receiver should send at least one ack for each window of data from the sender. Otherwise, it could cause the sender to stall until the receiver's delayed-ack timer (usually set at 200 ms) causes the resumption of the transfer.

4.2.2 Ack Filtering (AF)

The ACC mechanism described above modifies the TCP stack at the receiver in order to decrease the frequency of acks on the constrained reverse link. Ack filtering, based on an idea suggested by Karn [15], is a gateway-based technique that decreases the number of TCP acks sent over the constrained channel by taking advantage of the fact tha TCP acks are cumulative.

When an ack from the receiver is about to be enqueued, the router (o the end-host's routing layer, if the host is directly connected to the constrained link) traverses its queue to check if any previous acks belonging to the same connection are already in the queue. It then removes some fraction (or possibly all) of them, depending on how full the queue is. The removal of these "redundant" acks frees up space for other data and ack packets. The policy that the filter uses to drop packets is configurable and can either be deterministic or random (similar to a random-drop gateway, but using the semantics of the items in the queue). There is no need for any per-connection state to be maintained at the router — all the information necessary to implement the drop policy is already implicitly present in the packets in the queue.

In the experiments reported here, AF deterministically clears out all preceding acks belonging to a connection whenever a new ack for the same connection with a larger cumulative ack value enters the queue.

4.3 Results

We conducted a set of experiments, each involving a 50-second long transfer in the forward direction. There was no traffic in the reverse direction other than the acks for the forward transfer. Table 1 summarizes the throughputs obtained for three protocol configurations — regular TCP Reno, Reno with ACC and Reno with AF — with different types of return channels. With both ACC and AF, we include the sender adaptation techniques and perform traffic shaping to smooth out the spacing of acknowledgments.

The socket buffer size at the sender and receiver was set to 100 KB and each data packet was 1000 bytes long. The buffer size at each router was set to 10 packets. The ack size was set to 6 bytes and 40 bytes, respectively, with and without header compression.

The main observation here is that since the transfers are long, the reverse buffer fills up quickly. Beyond that point, the reverse bottleneck link limits the transfer. The greater the normalized asymmetry ratio (called *k*) is, the more limited the transfer becomes. In the steady state, only one ack in *k* gets through on average, causing the sender to send out bursts of *k* packets. As long as *k* does not exceed the bottleneck buffer size in the forward direction (which is 10 packets in our topology), the increased burstiness of the sender does not lead to losses.

Reverse Channel Bandwidth	Reno	ACC	AF
9.6 Kbps	1.78	3.64	6.28
9.6 Kbps C	6.67	7.69	7.93
28.8 Kbps	4.35	7.58	9.49
28.8 Kbps C	9.78	9.77	9.88

Table 1. Throughputs (in Mbps) for a single one-way transfer in the forward direction. "C" indicates the use of SLIP header compression.

The factor *k* exceeds 10 for the cases of SLIP without header compression, which explains the poor throughput of TCP Reno in those cases (1.78 and 4.35 Mbps). The sender adaptation employed in conjunction with ACC and AF smooths out sender bursts, avoiding performance degradation in those cases.

For the 9.6 Kbps reverse channel with header compression, *k* is 6.25, which is less than 10. Still the throughput obtained with TCP Reno

---

[2] The gateway can also be configured to drop the selected packet (Random Early Drop), but we chose to mark it instead.

(6.67 Mbps) is worse than that for the other schemes. This happens because the reverse channel buffer gets filled with acks (totalling 10*6 = 60 bytes), which adds a significant delay (60*8/9.6 = 50 ms) to the connections round-trip time (RTT). The same effect also explains why the performance with ACC is somewhat worse than that with AF in both the 9.6 Kbps and 28.8 Kbps cases. The former only tries to ensure that the reverse channel queue does not get completely filled up. The latter ensures that there is not more than one ack per connection in the queue, which minimizes the effect of queuing on round-trip time.

To summarize, TCP Reno suffers performance degradation when k is large and there is significant queuing delay. ACC and AF alleviate these problems by decreasing the frequency of acks.

4.4 Summary of Other Results

4.4.1 Two-Way Traffic

The presence of two-way traffic aggravates the problems due to bandwidth asymmetry discussed in Section 4.1. The transfer in the reverse direction uses up part of the limited reverse channel bandwidth. In addition, large data packets of the reverse transfer could cause acks of the forward transfer to be queued behind them for significant periods of time, further worsening performance.

Our solution to this is to have priority scheduling of acks (*acks-first* scheduling) at the reverse channel router. This minimizes the amount of time for which the ack stream is stalled. Our results show that combination of ACC and acks-first scheduling results in good performance in both directions.

4.4.2 Latency and Media-Access Asymmetry

In packet radio networks such as Metricom's Ricochet network (*http://www.metricom.com*), there may not be asymmetry in bandwidth, but the significant and variable latency for sending each packet causes problems. The presence of interfering traffic, including TCP acks, has an adverse effect on TCP data throughput because of variable latencies and large variations in round-trip time estimates. Our simulation experiments demonstrate that using ack filtering to decrease the frequency of acks results in significant improvement in both throughput for individual transfers as well as fairness when multiple connections are simultaneously active.

In the next section, we focus on how to improve TCP performance in the context of Web traffic.

5. Interactions of TCP and Web Traffic[5]

In the paper, we analyze the way in which Web browsers use TCP connections based on extensive traffic traces obtained from a busy Web server (the official Web server of the 1996 Atlanta Olympic games). At the time of operation, this Web server was one of the busiest on the Internet, handling tens of millions of requests per day from hundreds of thousands of clients. We present a detailed analysis of TCP's loss recovery and congestion control behavior from the recorded transfers. The two most important results are that: (1) short web transfers lead to poor loss recovery performance and (2) the use of parallel connections leads to overaggressive use of the network. We then discuss techniques designed to solve these problems. To improve the loss recovery performance of short transfers, we present a new technique for TCP loss recovery. To improve the congestion control and loss recovery performance of multiple parallel connections, we present a new transport-level *integrated connection* approach to congestion control and loss recovery. Simulation and trace analysis results show that our enhanced loss recovery scheme could have eliminated 25% of all timeout events, and our integrated connection approach provides greater fairness and improved startup performance for parallel connections. Our solutions are more general than a specific protocol enhancement such as the use of persistent connections in P-HTTP [19] and HTTP/1.1 [12], and include techniques, such as improved TCP loss recovery, that are not addressed by these protocols. Although the full paper presents a detailed analysis of connection behavior, we only discuss here the solutions to the problems found in the analysis.

5.1 Improving Single Connection Performance

5.1.1 Enhanced TCP Loss Recovery

Our analysis found that over 50% of all retransmissions on the Olympic Web server happened after one or more coarse timeouts kept the link to the client idle for periods from hundreds of milliseconds to seconds. Our analysis showed two main reasons for the occurrence of these timeouts:

1. Fast retransmit followed by a timeout: The TCP Reno sender was unable to recover from multiple losses within the same transmission window. This situation can be recognized by the occurrence of a fast retransmission shortly before the coarse timeout.
2. Insufficient duplicate acknowledgments: Either the number of outstanding packets was too small, or most of the packets in the window were lost, preventing the sender from receiving enough acknowledgment information to trigger a retransmission.

The use of TCP selective acknowledgments (SACKs) has often been suggested as a technique to improve loss recovery and avoid timeouts unless there is glan.tclenuine persistent congestion in the network. However, SACKs can only provide information to help recovery from multiple losses in the same window. A detailed analysis of the trace data showed that out of the 422025 coarse timeouts over a 3 hour period, SACKs could have helped avoid at most 18713 (4.43%) of them. *In other words, current approaches to TCP Reno enhanced with SACKs do not really avoid most timeouts.* It is clear that an alternative technique is needed to recover from the bulk of these losses.

During the same 3 hour period, approximately 403312 (95.6%) coarse timeouts occurred as a result of insufficient acknowledgment information (i.e., an insufficient number of duplicate acknowledgments arrived to trigger a retransmission). Of these timeouts, no duplicate acknowledgments arrived at all for 70% of them. In these situations the network is most likely experiencing severe congestion. The best solution is for the sender to wait for a coarse timeout to occur before transmitting any packets. For the remaining timeouts (about 25% of them, in which at least a single duplicate acknowledgment arrives), we propose that a single new segment, with a sequence number higher than any outstanding packet, be sent when each duplicate acknowledgment arrives. When this packet arrives at the receiver, it will generate an additional duplicate acknowledgment. When this acknowledgment later arrives at the sender, it can be assured that the appropriate segment has been lost and can be retransmitted. We call this form of loss recovery *enhanced* or *"right-edge"* recovery.

5.1.2 Results

An ns [18] based simulation was performed to test the enhanced loss recovery algorithm. A significant amount of additional cross traffic

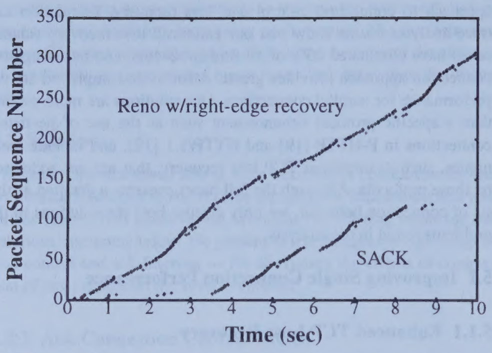


Figure 7. TCP with and without right-edge recovery.

was generated to force the transfers through the router to cope with frequent losses and small congestion windows. This was done to recreate situations that occur in the traces in a controlled fashion and not to simulate any existing or typical network topology. The simulation tests consisted of a single TCP transfer from node S to node R for a duration of 10 seconds. Each test used a different variant of TCP sender protocol on node S.

Figure 7 shows the simulated sequence plots for TCP with SACK and our enhancement to TCP-Reno with right-edge recovery. As indicated by the large gaps in the sequence plot, the SACK transfer experiences many more coarse timeouts than the transfer using right-edge recovery. These coarse timeouts are a result of frequent losses that generate two or fewer duplicate acknowledgments. These duplicates are not sufficient to trigger the fast retransmission algorithm. Therefore, in a traditional TCP implementation the loss is only retransmitted after a coarse timeout occurs. When the right-edge recovery algorithm is used, the sender transmits a new packet when these duplicate acknowledgments arrive. These new packets trigger the transmission of additional duplicate acknowledgments from receiver. These additional duplicates allow the sender to use fast retransmissions to recover from the losses. By eliminating the coarse timeouts, the sender using right-edge recovery performs the transfer more than twice as fast as a standard transmitter.

5.2 Improving Performance of Multiple TCP Connections

Current applications often use multiple logically separate data streams between a Web server and client. We showed in the analysis that the use of independent TCP connections for each transfer has a significant negative impact on network behavior. In this section, we present the design and implementation of a scheme that eliminates the negative impact of concurrent TCP connections and actually improves TCP loss recovery in such situations. This modified TCP uses integrated congestion control and loss recovery that exploits temporal parallelism and shares state across concurrent connections.

5.2.1 Integrated Congestion Control/Loss Recovery

The motivation behind kernel-level integrated congestion control and loss recovery is to allow applications to use a separate TCP connection for each transfer (just as they do today), but to avoid the problems that normally arise from multiple connections by making appropriate modifications to the network stack. We divide TCP functionality into two categories: that having to do with the reliable, ordered byte-stream abstraction of TCP and that pertaining to congestion control and data-driven loss recovery.

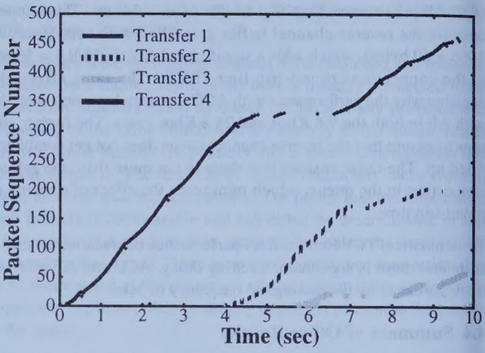


Figure 8. Four TCP-SACK transfers through a router with buffer size 3. Transfers start at 0, 2, 4 and 6 seconds.

By using a separate TCP connection for each transfer, an application can avoid unnecessary coupling between parallel transfers. Each connection provides a reliable, ordered byte-stream abstraction independently of the others. The flow control for each connection happens independently of the others, so the delivery of data to the receiving application also happens independently for each connection.

At the same time, congestion control is integrated across the TCP connections. There is a single congestion window for the set of TCP connections between a client and a server which determines how much total outstanding data the set of connections can have in the network. When a loss occurs on any of the connections, the combined congestion window is halved, thereby achieving the same effect as when a single, persistent TCP connection is used.

Data-driven loss recovery is also integrated across the set of TCP connections. When a packet is lost on one connection, the successful delivery of later packets on other connections allows the sender to reliably detect the packet loss without resorting to a timeout, thereby improving performance.

With such a network stack-based solution, existing unchanged applications can set up as many (possibly short) TCP connections as they wish without having an adverse effect on either congestion control or loss recovery.

5.2.2 Simulation Results: One Client Host Case

In this section, we describe results from an ns simulation designed to examine integrated congestion control and loss recovery across simultaneous TCP connections.

In the first test, the router buffers size was set to 3 packets. This is small enough to force transfers to have small congestion windows and experience frequent losses. Once again, the topology and parameters were chosen to recreate situations that frequently occur in our traces but not necessarily to mimic an actual network. In this test, the transmitting node performs 4 TCP transfers to the receiver. The transfers start at 0, 2, 4 and 6 seconds and all end at 10 seconds. The actual choices of the values 0, 2, 4, and 6 are not important, just that the start times of each connection are slightly shifted in time.

Figure 8 shows the sequence plot for the test using a selective acknowledgment based transmitter. It shows that typically only one

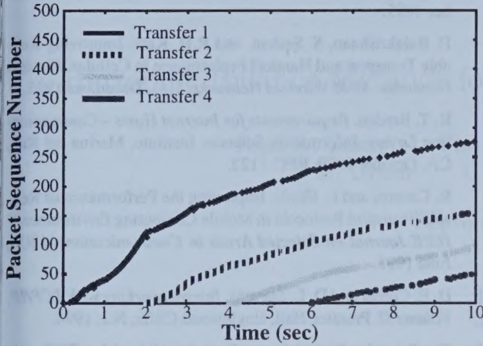


Figure 9. Four TCP-INT transfers through a router with buffer size 3. Transfers start at 0, 2, 4 and 6 seconds.

connection performs satisfactorily at any one time. For example, at time 2 seconds, the connection starting up experiences several early losses and is forced to recover them via coarse timeouts. In fact, this connection does not send a significant amount of data until 4 seconds later (at time 6 sec). Similarly, the first connection experiences many losses at time 4 seconds and spends several seconds recovering from these losses. Over the 10 second period, the connection starting at time 2 sec. and time 6 sec. account for a minuscule (<10%) fraction of the total bytes transferred. Although the combination of the four connections transmit data at a reasonable rate, the bandwidth is clearly not evenly split among the active connections. In addition, the numerous coarse timeouts make the performance of individual connections unpredictable. From an application's point of view, such unpredictability is undesirable because it may happen that connections carrying critical data get slowed down while others carrying less important data do better.

Figure 9 shows the sequence plot for the same test with the senders using TCP-Reno with integrated congestion control and loss recovery (henceforth called TCP-INT). Integrated loss recovery helps this TCP variant avoid resorting to coarse timeouts. Integrated congestion control allows the different connections to each obtain an equal share of the total bandwidth. Although the total number of bytes transferred here is actually slightly less than in the case with the selective acknowledgment protocol, the performance of the transfers is much more consistent and predictable. Also note that since the slow-start window growth happens in an integrated manner for the ensemble of connections, later connections can build on the window size attained by earlier connections rather than initiate slow start with an initial window size of one segment (which would happen with unmodified TCP). This can greatly help reduce the connection start-up latency for Web clients with short overlapping parallel connections.

5.2.3 Simulation Results: Multiple Client Hosts Case

We now investigate how the bottleneck link bandwidth is shared by connections initiated from more than one host. Our test uses the network topology shown in Figure 10. At time 0, a single TCP transfer is started from node S to each of nodes R1 and R2. Some time later (in this case, fifteen seconds), a second transfer is started between node S and R2. In addition, several cross traffic connection are maintained between nodes CT1 and CT2. This is intended to make the simulation more realistic.

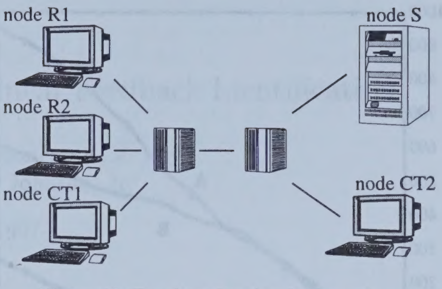


Figure 10. Topology for simulation tests.

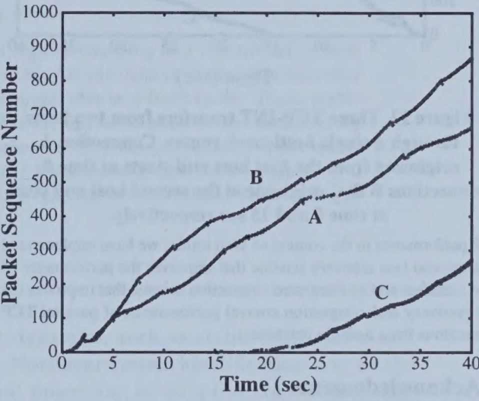


Figure 11. Three TCP-SACK transfers from two hosts through a single bottleneck router. Connection A originates from the first host and starts at time 0. Connections B & C originate at the second host and start at time 0 and 15 sec respectively.

Figure 11 shows the sequence plot for a test where node S uses TCP with selective acknowledgments. This graph shows that each connection receives approximately the same share of the bottleneck link bandwidth. As a result, node R2 receives approximately twice the bandwidth of node R1 after the second connection starts up. This is because each *connection* independently performs the same congestion control algorithms and are each equally likely to experience congestion. Figure 12 shows the same test using TCP-INT. After the second connection from node R2 starts, each of the transfers from R2 receive approximately half the bandwidth of the transfer on node R1. This is because each *host* independently performs the same congestion control algorithm. Therefore, each host receives an equal portion of the bottleneck link.

6. Conclusions

In this paper, we have presented improvements to TCP to improve performance in the context of wireless networks, asymmetric networks, and Web traffic. To improve the performance of TCP over wireless networks, we have implemented a TCP-aware link layer protocol called the snoop protocol that isolates wired senders from the lossy characteristics of a wireless link. To improve TCP performance over asymmetric networks, we have implemented techniques that reduce the effects of bandwidth asymmetry, when the forward path of a link has a higher link bandwidth than the reverse path. To improve

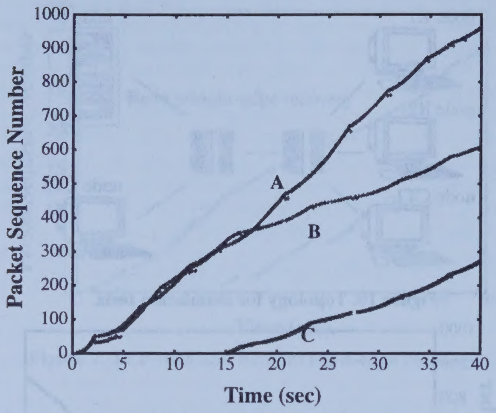


Figure 12. Three TCP-INT transfers from two hosts through a single bottleneck router. Connection A originates from the first host and starts at time 0. Connections B & C originate at the second host and start at time 0 and 15 sec respectively.

TCP performance in the context of Web traffic, we have implemented an improved loss recovery scheme that improves the performance of short transfers and an integrated connection scheme that improves the loss recovery and congestion control performance of parallel TCP connections from a single machine.

7. Acknowledgments

This work is supported by DARPA contract DAAB07-95-C-D154 and grants from the California MICRO Program, Hughes Aircraft Corporation, and IBM. Hari is partially supported by a Research Grant from the Okawa Foundation, and Mark is supported by an IBM fellowship.

8. References

[1] E. Amir, H. Balakrishnan, S. Seshan, and R. H. Katz. Efficient TCP over Networks with Wireless Links. In *Proc. Fifth IEEE Workshop on Hot Topics in Operating Systems*, May 1995.

[2] A. Bakre and B. R. Badrinath. I-TCP: Indirect TCP for Mobile Hosts. In *Proc. 15th International Conf. on Distributed Computing Systems (ICDCS)*, May 1995.

[3] H. Balakrishnan, V. N. Padmanabhan, and R.H. Katz. The Effects of Asymmetry on TCP Performance. In *Proc. ACM MOBICOM '97*, September 1997.

[4] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R.H. Katz. A Comparison of Mechanisms for Improving TCP Performance over Wireless Links. In *Proc. ACM SIGCOMM '96*, August 1996.

[5] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, M. Stemm, and R.H. Katz. TCP Behavior of a Busy Web Server: Analysis and Improvements. Technical report, University of California at Berkeley, 1997.

[6] H. Balakrishnan, S. Seshan, E. Amir, and R.H. Katz. Improving TCP/IP Performance over Wireless Networks. In *Proc. 1st ACM Conf. on Mobile Computing and Networking*, November 1995.

[7] H. Balakrishnan, S. Seshan, and R.H. Katz. Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks. *ACM Wireless Networks*, 1(4), December 1995.

[8] R. T. Braden. *Requirements for Internet Hosts – Communication Layers*. Information Sciences Institute, Marina del Rey, CA, October 1989. RFC-1122.

[9] R. Caceres and L. Iftode. Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments. *IEEE Journal on Selected Areas in Communications*, 13(5), June 1995.

[10] D. E. Comer and D. L. Stevens. *Internetworking with TCP/IP, Volume II*. Prentice Hall, Englewood Cliffs, N.J, 1994.

[11] The Daedalus Project Home Page. http://daedalus.CS.Berkeley.EDU/, 1995.

[12] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. *Hypertext Transfer Protocol – HTTP/1.1*, Jan 1997. RFC-2068.

[13] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, August 1993.

[14] V. Jacobson. Congestion Avoidance and Control. In *Proc. ACM SIGCOMM 88*, August 1988.

[15] P. Karn. Dropping TCP acks. Mail to the end-to-end mailing list, February 1996.

[16] P. Karn and C. Partridge. Improving Round-Trip Time Estimates in Reliable Transport Protocols. *ACM Transactions on Computer Systems*, 9(4):364–373, November 1991.

[17] S. McCanne and V. Jacobson. The BSD Packet Filter: A New Architecture for User-Level Packet Capture. In *Proc. Winter '93 USENIX Conference*, San Diego, CA, January 1993.

[18] ns: UCB/LBNL Network Simulator. http://www-mash.cs.berkeley.edu/ns/, 1997.

[19] V. N. Padmanabhan and J. C. Mogul. Improving HTTP Latency. In *Proc. Second International WWW Conference*, October 1994.

[20] W. R. Stevens. *TCP/IP Illustrated, Volume 1*. Addison-Wesley, Reading, MA, Nov 1994.

[21] AT&T WaveLAN: PC/AT Card Installation and Operation. AT&T manual, 1994.

[22] R. Yavatkar and N. Bhagwat. Improving End-to-End Performance of TCP over Mobile Internetworks. In *Mobile 94 Workshop on Mobile Computing Systems and Applications*, December 1994.

[23] L. Zhang, S. Shenker, and D. D. Clark. Observations and Dynamics of a Congestion Control Algorithm: The Effects of Two-Way Traffic. In *Proc. ACM SIGCOMM '91*, pages 133–147, 1991.

Harmonic Balance Methods for Nonlinear Feedback Identification

Raúl A. Casas* Clas A. Jacobson[†] G.J. Rey[†]
M. Green[‡] C. Richard Johnson, Jr.*

September 28, 1997

Abstract

Physical processes with complex dynamics often require modeling and identification from experimental data. This work proposes various methods for identifying static nonlinearities in a closed-loop configuration when the overall system operates in a limit cycle. These methods are based on harmonic balance analysis such that the approximation to the nonlinearity is not only accurate, but also admits limit cycle behavior. We discuss identification methods for two scenarios where the available data is (1) harmonically rich time series from a single experiment, or (2) estimated amplitude and frequency of oscillations from multiple experiments.

1 Introduction

Physical processes which exhibit complex nonlinear dynamics, such as stable limit cycles and bursting, require careful modeling and identification. Nonlinear system identification may be the crucial initial step towards solving an adaptive signal processing or adaptive control problem. For example, speech synthesis based on chaotic dynamics often sounds more natural than speech generated by linear oscillators. Similarly, combustion chambers which sustain stable pressure oscillations without external forcing are well modeled with nonlinear dynamics. Given a sufficiently accurate model, adaptive control for suppressing these oscillations may be designed and tested. In general, identification of some nonlinear dynamical systems can be divided into recovery of linear (G) and nonlinear (N) portions as shown in Figure 1 [Mees Book 81]. This work focuses on identification of the nonlinearity from experimental data when the linear system is known and the process is operating in a limit cycle.

The *information* used by the methods proposed in this paper are a combination of the linear system transfer function (G), and measured output data (y), precluding the need for feedback signal (u) measurements. These algorithms are based on harmonic balance and describing function analyses which are usually applied to the prediction of amplitude and frequency of limit cycles when the entire system is known (*i.e.* both G and N are available). They are extremely well suited for our problem, since one of the benefits of knowing the nonlinearity is the capability of predicting amplitude and frequency of oscillations, as well as determining their stability. For applications where this is the main underlying motivation, an accurate estimate of the nonlinearity is not the ultimate goal. Instead, one may rather think of finding a good estimate of the describing function of the nonlinearity, which is then used to make these predictions.

---

*School of Electrical Engineering, Frank H.T. Rhodes Hall, Cornell University, Ithaca, NY 14853, (607) 254-8819, {raulc,johnson}@ee.cornell.edu, supported by NSF Grant ECS-9528363 and United Technologies Research Center

[†]United Technologies Research Center, MS15, 411 Silver Lane, East Hartford, CT 06108, (860) 610-7652, {jacobsca,reygj}@utrc.utc.com, supported in part by DARPA Contract MDA 972-95-C009

[‡]Department of Engineering, Faculty of Engineering & Information Technology, Australian National University, Canberra, ACT 0200, Australia, (616) 249-2739 Michael.Green@anu.edu.au

Two algorithms derived from such concepts are introduced. The first algorithm, based on harmonic balance yields an estimate of the nonlinearity from a single experiment, where harmonically rich output data is available. When the harmonic content of the data is poor, a second method based on describing function analysis is more effective. This algorithm uses only estimates of the amplitudes of oscillations, but requires measurements from multiple, different experiments. This scheme immediately provides an estimate of the describing function of the nonlinearity, which can be used, as previously mentioned, for prediction of amplitudes and frequencies of oscillations.

The assumption that the linear system be known is strong and should be discussed. Knowledge of the linear dynamics can frequently be directly derived from physical laws, as for example in the bowing of a violin string, where a free body diagram provides sufficient information. More complex systems, however, may require identification from input/output data, in this case closed-loop data consisting of the system output (y) and feedback (u), with a zero reference signal. Linear system identification in closed-loop has been a popular research area since the 1970's [Gustavsson Auto 77, Pintelon TAC 94]. The majority of the work deals with closed-loop feedback from controllers, and most of the time it involves linear, rather than nonlinear feedback. Some results exist for the nonlinear feedback case, such as in [Ljung TAC 78] for *prediction error methods*, but it is not clear whether they apply to the limit cycle scenario. It should be noted that conventional linear system identification in closed-loop requires measurements of both output and feedback signals, while we are not assuming the availability of a measured feedback signal.

The paper is organized as follows: a brief discussion of the harmonic balance equations and describing function analysis leads to a separate derivation and discussion of each algorithm, along with error analyses based on *information based complexity* (IBC)[Traub Book 88] concepts. IBC offers a structured approach for the formulation of approximation problems. A case study implementing both algorithms and enumerating relevant observations concludes the paper.

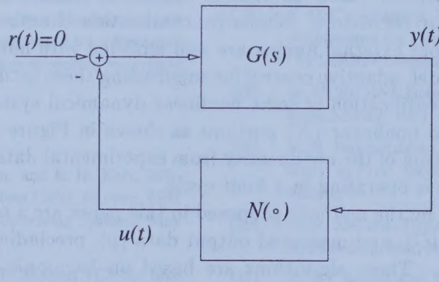


Figure 1: Nonlinear feedback limit cycle system.

2 Background

2.1 Harmonic Balance

From the system in Figure 1, using operator notation, the output y is given by

$$y = -G \cdot N \cdot y. \tag{1}$$

Here, the static nonlinearity $N \in \mathcal{N}$, $N : \mathbb{R} \mapsto \mathbb{R}$ is a time-invariant, memoryless, single-input single-output nonlinear function, while G is a linear delay-differential operator with transfer function $G(s)$.

When operating in a limit cycle, the signals in the closed-loop must necessarily be periodic. Define a period T (*i.e.* $y(t+T) = y(t)$) and frequency $\omega \equiv \frac{2\pi}{T}$. Then, y and u can be represented by Fourier series

$$y(t) = \sum_{k=-\infty}^{\infty} a_k e^{jk\omega t}, \; u(t) = N(y(t)) = \sum_{k=-\infty}^{\infty} c_k e^{jk\omega t}.$$

Note that G linearly scales and delays each frequency component so that

$$y(t) = \sum_{k=-\infty}^{\infty} a_k e^{jk\omega t} = -G \cdot u(t) = -\sum_{k=-\infty}^{\infty} c_k G(jk\omega) e^{jk\omega t}$$

and thus for all $k \in \mathbb{Z}$, we must then have what are known as the harmonic balance equations

$$a_k + G(jk\omega)c_k = 0. \tag{2}$$

2.2 Describing Function Analysis

When y consists mainly of one dominant harmonic (*i.e.* the term in the Fourier series with largest magnitude), the describing function (DF) of the nonlinearity may be used to predict the approximate amplitudes and frequencies of possible limit cycles.

We briefly outline the basic theory for the case where N has odd symmetry

$$N(x) \;=\; -N(-x), \;\; \forall x \neq 0$$

and lies in a bounded sector

$$\alpha(x_2 - x_1) \leq N(x_2) - N(x_1) \leq \beta(x_2 - x_1), \;\; \forall x_2, x_1 \in \mathbb{R}, \;\; \alpha, \beta > 0.$$

Write y as the sum of a dominant frequency component and other smaller terms

$$y(t) \;=\; A \sin(\bar{\omega} t) + y_h(t),$$

where $\bar{\omega} = \bar{k}\omega$ for some $\bar{k} \in \mathbb{Z}$ (note that we have translated time (w.l.o.g.) so that the main harmonic has no phase offset). Then,

$$\left| \frac{1}{G(j\bar{\omega})} + \Psi(A) \right| \leq \sigma(\bar{\omega}) \tag{3}$$

where Ψ is the describing function of N given by

$$\Psi(a) \;=\; \frac{2}{\pi a} \int_0^{\pi} N(a \sin\theta) \sin\theta d\theta \tag{4}$$

and $\sigma(\omega)$ is a finite bound on the error in (3) (refer to the discussion in [Khalil Book 92], p.364). The DF method makes the approximation

$$G(j\bar{\omega}) \;\approx\; -\frac{1}{\Psi(A)}$$

which says that the amplitudes A and frequencies $\bar{\omega}$ for (possible) limit cycles of the system in Figure 1 are approximately given by the intersection of the Nyquist diagram of G and the locus of $-\frac{1}{\Psi(a)}$. Since $\Psi(a)$ is a real quantity, this intersection occurs on the negative real axis of the complex plane. For more details, see [Khalil Book 92, Gelb Book 68, Mees Book 81].

3 Nonlinearity ID Algorithms

3.1 Single Experiment

The first experimental scenario under consideration is the case when only one output data record is available. This time series is harmonically rich, containing several strong harmonics, or Fourier components. This wealth of information from a single experiment along with good estimates of the transfer function G is used to recover the nonlinearity N.

From (2) $c_k = -\frac{a_k}{G(jk\omega)}$ if $G(jk\omega) \neq 0$. Now, u, and consequently c_k, are nonlinear functions of y. By parameterizing this nonlinear map appropriately we can derive linear equations in these parameters using each computed Fourier component. One way to do this consists of approximately reconstructing N through a linear combination of basis functions L_i, which in the best of all worlds would not only span the space where N lies but also render a set of well-conditioned equations.

Pick a basis of functions $\{L_i\}_{1\leq i\leq M}$, $L_i \in \mathcal{N}$ and parameterize the approximation with parameters $\alpha_i \in \mathbb{R}$ as

$$\hat{N}(x) \;=\; \sum_{i=1}^{M} \alpha_i L_i(x).$$

This parameterization yields equations which are linear in α_i

$$c_k + \tilde{c}_k = \frac{1}{T}\int_0^T \hat{N}(y(\tau))e^{jk\omega\tau}d\tau = \sum_{i=1}^{M} \alpha_i \underbrace{\frac{1}{T}\int_0^T L_i(y(\tau))e^{jk\omega\tau}d\tau}_{W_{ki}} = -\frac{a_k}{G(jk\omega)}$$

and can be solved via least squares ($\tilde{c}_k$ reflects the fact that if the parameterization is approximate then the equations above will not be equalities). We call this the Harmonic Balance Nonlinearity Identification (HB-NID) algorithm.

Remarks

- A straightforward parameterization uses polynomials $L_i(x) = x^{i-1}$ so that $N(x) = \sum_i \alpha_i x^{i-1}$. Polynomials may be well suited for local approximations and small amplitude oscillations, but their unbounded rate of growth may not admit large amplitude limit cycles.
- Note that the equations above project $L_i(y(t))$ onto the Fourier basis over one period; for averaging purposes several periods could be used.
- Certain harmonics crucial for identification may not be available when $G(jk\omega) = 0$, as is the case for the DC component when G has differentiators.
- Adaptive implementation of this algorithm for applications requiring online real-time identification is possible by defining the mean square error (MSE) $\lim_{T\to\infty} \frac{1}{T}\int_0^T \varepsilon^2(\tau)d\tau$ where

  $$\varepsilon(t) \;=\; \sum_k \left(\sum_i \alpha_i(t) W_{ki}(t) + \frac{a_k}{G(jk\omega)} \right) e^{jk\omega t}$$

  is the instantaneous error and $W_{ki}(t) = \frac{1}{T}\int_{t-T}^{t} L_i(y(\tau))e^{jk\omega\tau}d\tau$. The MSE may be minimized over the parameters α_i via a stochastic gradient descent scheme such as the least mean square algorithm (LMS).

Error Analysis

The available *information* in this problem is a finite set of pairs of Fourier coefficients and transfer function gains $I_N = \{(a_k, G(jk\omega))\}_{k\in K}$, for some finite set $K \subset \mathbb{Z}$ of harmonics corresponding to closed-loop limit cycle data using nonlinear feedback $N \in \mathcal{N}$. We fix, w.l.o.g., the power of u to be unity. Define the *problem set*, the set of functions we wish to approximate, to be $\mathcal{N} = (C(X,\mathbb{R}), \|\cdot\|_\infty)$ where $X \subset \mathbb{R}$ is the interval over which the approximation is done. Since N operates on y, the interval X should be chosen to match the range of values of y (*i.e.* if the oscillations y have amplitude A, then $X = [-A, A]$). The set of all possible nonlinearities that satisfy the available information, or the set of *indistinguishable problem elements*, is

$$I_N^{-1} = \{f \in \mathcal{N} : I_f = I_N\}$$

We wish to compute the *local radius of information*, which measures the the radius of the smallest ball containing I_N^{-1}

$$r(I_N) = \inf_{f\in N} \sup_{\tilde{f}\in I_N^{-1}} \|f - \tilde{f}\|.$$

Let $p_K = \sum_{k\in K} |c_k|^2$ be the power of successive Fourier approximations of u, so that $p_\infty = 1$. Now,

$$\begin{aligned} I_N^{-1} &= \{f \in \mathcal{N} : f(y(t)) = \sum_{k=-\infty}^{\infty} b_k e^{jk\omega t} \\ & \qquad b_k = -\frac{a_k}{G(jk\omega)}, k \in K, \sum_{k=-\infty}^{\infty} |b_k|^2 = 1\} \\ &= \{f \in \mathcal{N} : \sum_{k\notin K} |b_k|^2 = 1 - p_K\}. \end{aligned}$$

Thus, I_N^{-1} is an ℓ_2-norm *sphere*, and the smallest ℓ_∞-norm ball around it has radius $r(I_N) = \sqrt{1 - p_K}$ (see Figure 2). Observe that up to this point the error analysis is independent of the algorithm used for approximation. The radius of information is a quantity fundamental to the problem, which indicates that to reduce uncertainty, more harmonics should be used.

The *worst case local error* for an algorithm which yields an approximation $\hat{N}$ from information I_N is defined as

$$e(\hat{N}, I_N) = \sup_{\tilde{f}\in I_N^{-1}} \|\tilde{f} - \hat{N}\|,$$

Thus, for a given choice of basis defining a *flavor* of HB-NID algorithm, we could derive an expression for the worst case error. This worst case error will always be bounded below by the radius of information, *i.e.* $e(\hat{N}, I_N) \geq r(I_N)$. When equality is possible, we say the algorithm is *optimal.*

Notice that if the algorithm always provides approximations which lie in I_N^{-1} it will necessarily be optimal. Thus, any choice of basis for HB-NID which always provides K independent equations (thus requiring at least K parameters/basis functions) is potentially optimal. An additional requirement not resolved by HB-NID, however, is that $\hat{N}(y(t))$ have unity power, and therefore, in general, it is not an optimal algorithm.

3.2 Multiple Experiments

In the second experimental scenario we look at the alternative case where multiple experiments are conducted but only produce measurements of the base frequency and amplitude of oscillations. The DF method is perfectly suited for developing an identification algorithm for this setting.

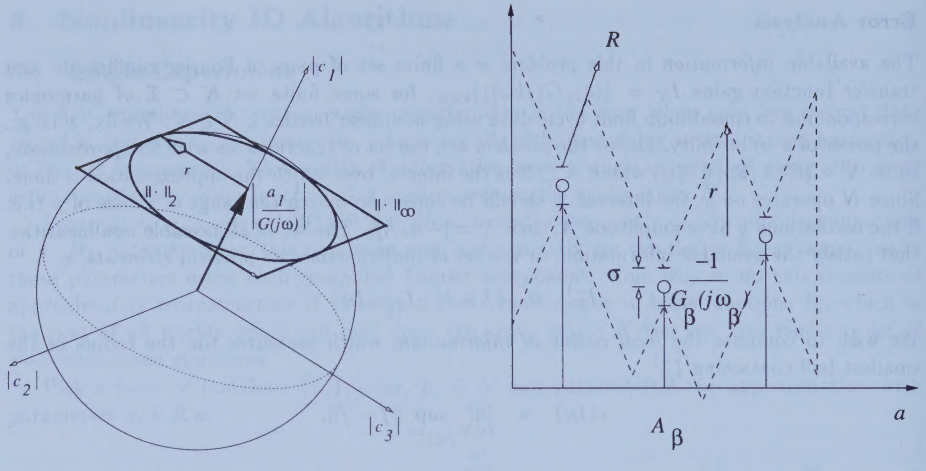


Figure 2: (left) Geometric interpretation of radius of information for HB-NID and (right) for bowtie regions for describing functions of DF-NID.

Let each experiment $\beta, 1 \leq \beta \leq P$ use a linear portion G_β. In each experiment the linearity may change by varying its gain or time delay, for example. Each experiment gives estimates A_β of amplitude and ω_β of frequency of the oscillations.

Using the triplet $(G_\beta, A_\beta, \omega_\beta)$, the DF approximation is

$$\Psi(A_\beta) \approx -\frac{1}{G_\beta(j\omega_\beta)}$$

(with error bound $\sigma_\beta(\omega_\beta)$). A parameterization of N with basis $\{L_i\}_{1\leq i\leq M}$ leads to the Describing Function Nonlinearity Identification (DF-NID) algorithm

$$\Psi(A_\beta) \approx \sum_{i=0}^{M} \alpha_i \underbrace{\frac{2}{\pi A_\beta}\int_0^\pi L_i(A_\beta \sin\theta)\sin\theta d\theta}_{V_{\beta i}} = \underbrace{-\frac{1}{G_\beta(j\omega_\beta)}}_{\gamma_\beta}, \quad 1 \leq \beta \leq P. \tag{5}$$

Remarks

- Once more the need for accurate estimates of the linear system(s) G_β are required. Note that as the error bound $\sigma(\omega)$ of DF approximation increases the estimates of N degrade in quality. This will be a crucial issue for the selection of the basis functions L_i, since the conditioning of the set of equations to be solved depends on these functions.
- Because the DF analysis described above only applies to nonlinearities with odd symmetry, DF-NID will only identify nonlinearities up to an even function.

Error Analysis

For this algorithm we will study the error in the estimate of the describing function, rather than in the nonlinearity approximation. The *information* consists of the amplitude and frequency of oscillations and the complex gain of the linear system at that frequency for each experiment, *i.e.* the set of triplets $I_\Psi = \{(G_\beta, A_\beta, \omega_\beta)\}_{1 \le \beta \le P}$, some $P < \infty$. The problem set is the class of describing functions with bounded rate

$$V = \{f \in C^1(\mathbb{R}, \mathbb{R}) : |f'(a)| \le R \;\; \forall a \in \mathbb{R}\}.$$

The restriction on the derivative is basically a result of bounding the derivative of nonlinearities, and will aid in providing a set with finite radius of information. The set of *indistinguishable problem elements* for this case is

$$I_\Psi^{-1} = \{\psi \in V : I_\psi = I_\Psi\} = \{\psi \in V : \psi \in \cap_\beta \mathcal{B}_\beta\}$$

where $\mathcal{B}_\beta$ are the *bowtie* regions

$$\mathcal{B}_\beta = \left\{\psi \in V : \left|\psi(a) + \frac{1}{G_\beta(\omega_\beta)}\right| \le \sigma_\beta(\omega_\beta) + R|a - A_\beta|\right\}$$

shown in Figure 2. Let $\Delta_\beta = A_\beta - A_{\beta-1}$, and assume that $\sigma_\beta(\omega_\beta) < \sigma$ for some $\sigma \in \mathbb{R}$, for all β. Analysis of the bowtie regions shows that the local radius of information given experimental data I_Ψ is

$$r(I_\Psi) = \sigma + R \max_\beta \Delta_\beta.$$

Finally, we define the local error $e(\hat{\Psi}, I_\Psi) = \sup_{\tilde{f} \in I_\Psi^{-1}} \|\tilde{f} - \hat{\Psi}\|$, which once again satisfies $e(\hat{\Psi}, I_\Psi) \ge r(I_\Psi)$. This analysis says that in order to reduce the lower bound on the worst case error the intervals between amplitudes measured by experiments (*i.e.* Δ_β) must be reduced. Inherent limitations introduced by the "difficulty" of the problem, such as the DF error bound σ and the nonlinearity rate of growth bound R, are in general unavoidable.

4 Case Studies

This section works out two examples illustrating the use of both algorithms and pointing out important observations regarding their implementation.

4.1 Odd Symmetric NID

We begin with a second-order linear system with delay

$$G(s) = \gamma e^{-s\delta} \frac{\omega_n^2}{s^2 + 2\xi\omega_n + \omega_n^2} \qquad \begin{array}{c|c} \gamma & 3 \\ \xi & 0.18 \\ \omega_n & 2\pi \\ \delta & 0.1 \end{array}$$

and nonlinear feedback given by the sigmoidal function

$$N(x) = \frac{e^x - 1}{e^x + 1}.$$

Because DF-NID will only identify nonlinearities up to an even symmetric function, we opt for this odd symmetric nonlinearity to show successful recovery.

The closed-loop system in limit-cycle produces the periodic signals shown in Figure 3. The DF argument illustrated in Figure 3 correctly predicts oscillations with approximate amplitude 6.0 and period 0.83.

As mentioned before, the nonlinearity identification objective is not only to obtain a good approximation but also to have an approximation which admits stable limit-cycles and accurately predicts the amplitude of the oscillations. Along with *a priori* information about the

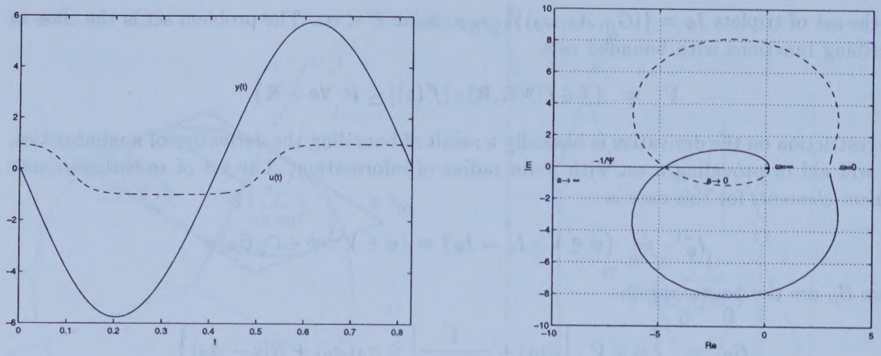


Figure 3: (left) Input u and output y waveforms, (right) DF analysis: intersection of Nyquist diagram of $G(\omega)$ and DF locus of $-\frac{1}{\Psi(a)}$.

nonlinearity, these issues are crucial for the selection of the basis functions L_i. The clamping effect on the extreme ranges of the waveform u (if available) as a function of y is good evidence that the nonlinearity exhibits saturation. By imposing saturation on the basis functions, their linear combination will also exhibit saturation. Note that one could still arrive at this requirement by observing that saturation conveniently provides boundedness of the signals in closed-loop.

Further inspection of u and y indicates odd symmetry of the nonlinearity, further restricting L_i to be odd symmetric. We chose a straightforward piecewise linear approximation with saturated basis functions

$$L_i(x) \;=\; \begin{cases} x & x \le x_i \\ x_i \operatorname{sgn}(x) & x > x_i \end{cases} \tag{6}$$

For HB-NID, the output lies in $X = [-A, A]$, $A \approx 6$, and we select $x_i = 1, 2, 3, 4$. The algorithm gives the approximation to N shown in Figure 4. For DF-NID, we set the breaks x_i to match the measured amplitudes A_β. Figure 4 compares the DF of N and of the approximation given by DF-NID.

4.2 Combustion Chamber Model

We now investigate nonlinear feedback in a model of a combustion chamber with unsteady pressure oscillations [Murray ACC 98]. A second order model for the linear acoustics of the combustion chamber is

$$G(s) = \gamma e^{-s\delta} \frac{s^2\omega_n^2}{s^2 + 2\xi\omega_n + \omega_n^2} \qquad \begin{array}{|cc} \gamma & 0.0197 \\ \xi & 0.18 \\ \omega_n & 2\pi \\ \delta & 0.5 \end{array}$$

and a crude model for the nonlinear relation between pressure and heat release is

$$N(x) \;=\; \log(e^{x-x_0}+1), \;\; x_0 = -\frac{1}{2}.$$

For this example, a polynomial basis $\{x^{i-1}\}_{1\le i\le 6}$ gives good local approximations of N as shown in Figure 5. Outside the interval where data is available, however, the approximation

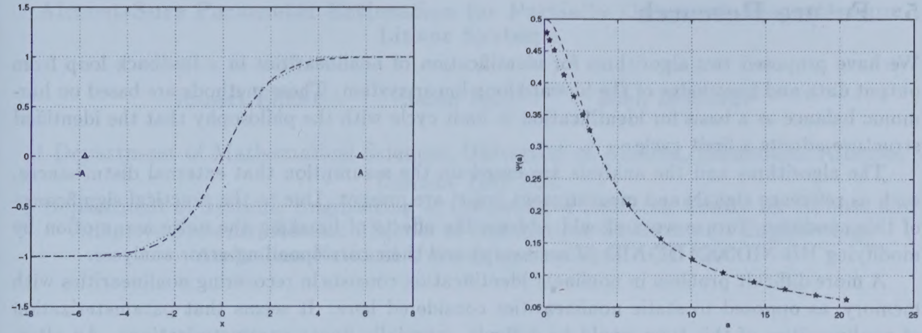


Figure 4: (left) Piecewise linear approximation of sigmoidal function using HB-NID. Dashed line is $N(x)$ and dotted line is the approximation. The Δ's delineate the interval $X = [-A, A]$ over which data is available. (right) Describing function of piecewise linear approximation of sigmoidal function using DF-NID. Dashed line is DF of $N(x)$ and dotted line is DF of approximation. The $\star$'s mark measured amplitudes of oscillations A_β for each experiment.

is poor, and in fact grows as x^5, a condition that goes against the philosophy of preserving qualitative characteristics of the dynamics. Observe in Figure 5 that the locus of the DF of the true nonlinearity covers a segment of the negative real axis, more explicitly, it is located between the (negative) inverse equivalent average gain of the nonlinearity for small oscillations at $-1/N'(0) \approx -1.6$, and large oscillations at $-2/N'(\infty) \approx -2$. Oscillations grow due to a large initial gain and die out for the small average gain, settling at some intermediate operating average gain, near the intersection with the Nyquist plot of the transfer function. The polynomial expansion/approximation of the nonlinearity has a DF with unbounded growth and a locus that approaches the origin for large amplitude oscillations. Hence, the qualitative behavior displayed by the DF of the true nonlinearity is not maintained. The polynomial expansion, however, has the advantage of allowing accessible perturbation/bifurcation analysis of the underlying nonlinear dynamical system at hand.

Notice that due to the double differentiator, DC bias information is lost as can be seen from input/output data in the same figure. In fact, the constant term coefficient (*i.e.* for $L_1(x) = 1$) is ill-computed, and is therefore manually adjusted in the figure for comparison purposes.

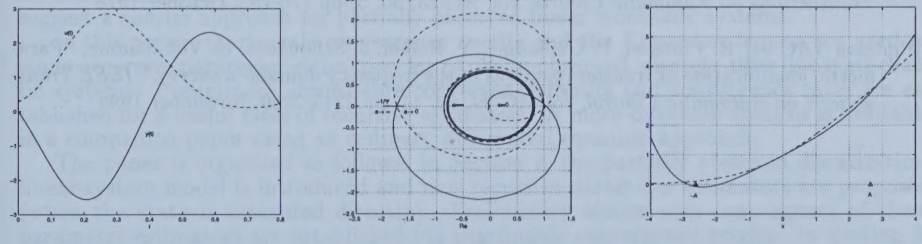


Figure 5: (left) Input u and output y waveforms, (middle) DF analysis: intersection of Nyquist diagram of $G(\omega)$ and DF locus of $-\frac{1}{\Psi(a)}$, (right) Actual heat release nonlinearity (dashed) and HB-NID polynomial approximation (solid). The Δ's delineate the interval $X = [-A, A]$ over which data is available.

5 Future Research

We have proposed two algorithms for identification of nonlinearities in a feedback loop from output data and knowledge of the forward loop linear system. These methods are based on harmonic balance as a basis for identification in limit cycle with the philosophy that the identified structure admits a limit cycle.

The algorithms and the analysis are based on the assumption that external disturbances, such as reference signals and measurement noise, are present. Due to the practical significance of this condition, future work should address the effects of breaking the noise assumption by modifying HB-NID and DF-NID (if necessary) and their corresponding error analyses.

A more difficult problem in nonlinear identification consists in recovering nonlinearities with memory, as opposed to static nonlinearities considered here. It seems that parameterization of nonlinearities of this type would be difficult, especially linear parameterizations. An alternative method for this task could potentially use a nonparametric scheme, where a *graph* of input/output trajectories traces the nonlinearity.

References

[Mees Book 81] A.I. Mees, *Dynamics of Feedback Systems*, John-Wiley & Sons, New York, 1981

[Gelb Book 68] A. Gelb, W.E. Vander Velde, *Multiple-Input Describing Functions and Nonlinear System Design*, McGraw-Hill, Inc., New York, 1968

[Khalil Book 92] H.K. Khalil, *Nonlinear Systems*, Macmillan Publishing Company, New York, NY, 1992

[Murray ACC 98] R.M. Murray, C.A. Jacobson, R.A. Casas, A.I. Khibnik, C.R. Johnson, Jr., R.R. Bitmead, A.A. Peracchio, W.M. Proscia, "System identification for limit cycling systems: a case study for combustion instabilities," *Submitted for publication in the proceedings of the 1998 American Control Conference*

[Traub Book 88] J.F. Traub, G.W. Wasilkowski, H. Wozniakowski, *Information-Based Complexity* Academic Press, Inc., San Diego, CA, 1988

[Gustavsson Auto 77] I. Gustavsson, L. Ljung, T. Söderström, "Identification of processes in closed loop - identifiability and accuracy aspects," *Automatica*, vol. 13, pp.59-75, 1977

[Ljung TAC 78] L. Ljung, "Convergence analysis of parametric identification methods," *IEEE Transactions on Automatic Control*, vol. AC-23, no. 5, pp.770-783, October 1978

[Pintelon TAC 94] R. Pintelon, P. Guillaume, Y. Rolain, J. Schoukens, H. Van hamme, "Parametric identification of transfer functions in the frequency domain- a survey," *IEEE Transactions on Automatic Control*, vol. AC-39, no. 11, pp.2245-2260, November 1994

Almost-Sure Parameter Estimation for Partially Observed Discrete-time Linear Systems

Robert Elliott† Jason Ford‡ John B. Moore‡

† Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1
‡ Department of Systems Engineering & Cooperative Research Centre for Robust and Adaptive Systems, Research School of Information Sciences and Engineering, Australian National University, Canberra ACT 0200.

Abstract

In this paper we discuss parameter estimators for fully and partially observed discrete-time linear stochastic systems. We propose parameter estimators that do not need direct state information but are based on estimates of functions of the state. We establish preliminary almost-sure convergence of parameter estimators based on conditional mean estimates using standard martingale convergence results and the Kronecker lemma. A more complete analysis is available in a companion paper using an ordinary differential equation approach. We also provide simulation studies which demonstrate the performance of these estimators.

1 Introduction

The identification of linear systems with measured input and output sequences has been studied extensively, see [1-7]. In fully observed linear systems the linear relationship between input and output signals ensures global convergence of parameter estimates, see [5, 6]. However, for partially observed systems, when functions of input signals or state variables need to be estimated from output measurements, no global convergence results for system identification are given in the existing literature. Off-line locally convergent algorithms based on the EM algorithm are available [9].

Recently, almost-surely convergent parameter estimators algorithms have been proposed for hidden Markov models in discrete and continuous time, [10, 11], and linear systems in continuous time [12]. In particular, the almost-surely convergent parameter estimators proposed for hidden Markov models in discrete time are based on estimates for functions of the Markov state [10]. These estimators and their convergence proofs suggest a similar approach for partially observed linear stochastic systems.

In this paper, martingale convergence results and the Kronecker lemma are used to study recursive parameter estimation for partially-observed, discrete-time linear stochastic systems. Preliminary almost-sure convergence results and convergence rates are established for a useful class of recursive estimators. A more complete analysis is available in a companion paper using an ordinary differential equation approach.

The paper is organised as follows: in Section 2, the partially observed discrete-time linear system model is introduced and new recursive parameter estimators are proposed (when the state is measured directly). Preliminary almost-sure convergence of these parameter estimators are established via martingale convergence results. In Section 3, estimation is discussed when the state is not measured directly, the relevant conditional mean filters are developed, parameter estimators proposed and almost-sure convergence results established. A new finite dimensional conditional mean filter is also presented. In Section 4 we present some simulation studies. Conclusions are given in Section 5.

2 Dynamics

Consider a probability space $(\Sigma, \mathcal{F}, P)$; suppose $\{x_\ell\}$, $\ell \in \mathcal{Z}^+$ is a discrete-time linear stochastic process, taking values in R^N, with dynamics: given by

$$x_{k+1} = Ax_k + Bw_{k+1}, \quad x_0 \in R^{N\times 1}. \tag{1}$$

Here $k \in \mathcal{Z}^+$, $A \in R^{N\times N}$, $B \in R^{N\times 1}$ and $\{w_\ell\}$, $\ell \in \mathcal{Z}^+$ is a sequence of independent and identically distributed $N(0,1)$ scalar random variables. Denote the probability density of the noise, Bw_ℓ, as ψ.

The state process x_ℓ, $\ell \in \mathcal{Z}^+$ is observed indirectly via the scalar observation process $\{y_\ell\}$, $\ell \in \mathcal{Z}^+$, given by

$$y_k = Cx_k + Dv_k. \tag{2}$$

Here for $k \in \mathcal{Z}^+$, $C \in R^{1\times N}$, $D \in R$ and $\{v_\ell\}$, $\ell \in \mathcal{Z}^+$ is a sequence of independent and identically distributed $N(0,1)$ scalar random variables. Denote the probability density of the noise, Dv_ℓ as ϕ. We assume that x_0, $\{v_\ell\}$ and $\{w_\ell\}$ are mutually independent.

Let $\mathcal{F}_k = \sigma\{x_0, x_1, \ldots, x_k\}$ and $\mathcal{Y}_k = \sigma\{y_0, y_1, \ldots, y_k\}$ denote the filtrations generated by x and y respectively. Also, let $\mathcal{G}_k = \sigma\{x_0, \ldots, x_k, y_0, \ldots, y_k\}$ denote the filtration generated by x and y.

Denote the model (1), (2) by

$$\lambda = \lambda(A, B, C, D, x_0) \tag{3}$$

2.1 Measure Change

Define the probability measure $\bar{P}$ such that the $\mathcal{G}_k$ restriction of the Rayon-Nikodym derivative of P with respect to $\bar{P}$ is:

$$\left.\frac{d\bar{P}}{dP}\right|_{\mathcal{G}_k} = \Gamma_k = \prod_{\ell=1}^{k} \gamma_\ell \quad \text{where } \gamma_\ell = \frac{\psi(Bw_\ell)\phi(Dv_\ell)}{\psi(x_\ell)\phi(y_\ell)} \tag{4}$$

Theorem 1 *Under the measure $\bar{P}$:*
1. $\{x_\ell\}$ and $\{y_\ell\}$ are independent random variable with densities ψ and ϕ, respectively. $\{x_\ell\}$ and $\{y_\ell\}$ are also independent of each other.
2. If f_ℓ is a $\mathcal{G}$ adapted sequence then the following abstract version of Bayes theorem holds:

$$E[f_\ell|\mathcal{Y}_k] = \frac{\bar{E}[\bar{\Gamma}_\ell f_\ell|\mathcal{Y}_k]}{\bar{E}[\bar{\Gamma}_\ell|\mathcal{Y}_\ell]} \tag{5}$$

where $\bar{E}[.]$ denotes expectation in $\bar{P}$ and $\bar{\Gamma}$ denotes the reverse measure change.

Proof: see [1].

Identity (5) enables us to obtain the conditional expectation $E[f_\ell|\mathcal{Y}_k]$ if we know the un-normalised conditional expectation $\bar{E}[\bar{\Gamma}_\ell f_\ell|\mathcal{Y}_k]$.

2.2 Parameter Estimation - Full Observations

In this subsection we assume that both $\{x_k\}$ and $\{y_k\}$ are fully observed. The results in this section for the full observation case are presented as a stepping stone to the more interesting and general results of Section 3.

From (1), by post-multiplication by x'_k and summing, we obtain

$$\sum_{i=1}^{k} x_i x'_{i-1} = A\sum_{i=1}^{k} x_{i-1}x'_{i-1} + B\sum_{i=1}^{k} w_i x'_{i-1}. \tag{6}$$

Here the prime $'$ denotes the transpose operation. Now consider the matrices

$$J_k = \sum_{i=1}^{k} x_i x'_{i-1} \qquad O_k = \sum_{i=1}^{k} x_{i-1} x'_{i-1} \qquad M_k = \sum_{i=1}^{k} w_i x'_{i-1}. \tag{7}$$

From (6) we see $J_k = AO_k + BM_k$. Assuming that x_k is measurable, a reasonable estimate for A is, therefore,

$$\hat{A}_k = J_k O_k^{-1} \tag{8}$$

when O_k^{-1} exists. The error in this estimate is $\hat{A}_k - A = BM_k O_k^{-1}$.

Similarly, from (2), by post-multiplication by x'_k and summing we obtain

$$\sum_{i=1}^{k} y_{i-1} x'_{i-1} = A \sum_{i=1}^{k} x_{i-1} x'_{i-1} + D \sum_{i=1}^{k} v_{i-1} x'_{i-1}. \tag{9}$$

Now consider the vectors

$$T_k = \sum_{i=1}^{k} y_{i-1} x'_{i-1} \quad V_k = \sum_{i=1}^{k} v_{i-1} x'_{i-1}. \tag{10}$$

Then (9) can be written $T_k = CO_k + DV_k$. A reasonable estimate for C is, therefore,

$$\hat{C}_k = T_k O_k^{-1} \tag{11}$$

when O_k^{-1} exists. The error in this estimate is $\hat{C}_k - C = DV_k O_k^{-1}$.

To obtain estimates for B and D we square element-wise (1) and (2), respectively, and summing we obtain

$$((b^1)^2, \ldots, (b^N)^2) \sum_{\ell=1}^{k} w_\ell^2 = \sum_{\ell=1}^{k} ((x_\ell^1)^2, \ldots, (x_\ell^N)^2) - \sum_{\ell=1}^{k} \left((\sum_{i=1}^{N} a_{1i} x_{\ell-1}^i)^2, \ldots, (\sum_{i=1}^{N} a_{Ni} x_{\ell-1}^i)^2 \right) \tag{12}$$

and

$$D^2 \sum_{\ell=1}^{k} v_\ell^2 = \sum_{\ell=1}^{k} y_\ell^2 - \sum_{\ell=1}^{k} \left((\sum_{i=1}^{N} C_i x_\ell^i)^2 \right) \tag{13}$$

where $B = (b^1, \ldots, b^N)'$. Equations (12) and (13) can be used to estimate B and D respectively as

$$\left((\hat{b}_k^1)^2, \ldots, (\hat{b}_k^N)^2\right) = k^{-1} \left(\sum_{\ell=1}^{k} ((x_\ell^1)^2, \ldots, (x_\ell^N)^2) - \sum_{\ell=1}^{k} \left((\sum_{i=1}^{N} a_{1i} x_{\ell-1}^i)^2, \ldots, (\sum_{i=1}^{N} a_{Ni} x_{\ell-1}^i)^2 \right) \right) \tag{14}$$

and

$$\hat{D}_k^2 = k^{-1} \left(\sum_{\ell=1}^{k} y_\ell^2 - \sum_{\ell=1}^{k} \left((\sum_{i=1}^{N} C_i x_\ell^i)^2 \right) \right) \tag{15}$$

We shall investigate the convergence to zero of the error in these estimators.

2.3 Almost-Sure Convergence

In this subsection we discuss the convergence of these estimators. Before proceeding to our convergence results we first state a stability condition for a linear system.

A time-invariant system with state transition matrix A is *stable* if the following condition holds

$$\sigma_{\max}(A) < 1 \tag{16}$$

where $\sigma_{\max}(A)$ is the largest magnitude of the eigenvalues of the matrix A.

The system (1),(2) is said to be *persistently excited* if:

$$\lim_{k\to\infty}\sup_{k}\left(\frac{1}{k}O_k\right)^{-1} \quad \text{exists.} \tag{17}$$

Now consider

$$R_k := \sum_{i=1}^{k} \rho(i)^{-\frac{1}{2}}\Delta M_i.$$

Here, the function $\rho(k), k \geq 0$, is positive, non-decreasing and such that $\lim_{k\to\infty}\sum_{i=0}^{k}\rho(i)^{-1} = \lambda < \infty$ and $\Delta M_i := M_k - M_{k-1} = w_i x'_{i-1}$. An example of $\rho(k)$ satisfying this requirement is $\rho(k) = \max(1, klnk(lnlnk)^{\alpha})$, for any $\alpha > 1$.

Lemma 1 *Suppose the system (1),(2) is stable such that (16) holds. Then R_k is a vector whose elements, R^i_k for $i = 1,\ldots,N$, are square integrable martingales so that $\lim_{k\to\infty} R^i_k = \lambda^i(\omega) < \infty$ exists a.s.*

Proof: R^i_k is a martingale because $E[R^i_k|\mathcal{F}_{k-1}] = E[\rho(k)\Delta M^i_k + R^i_{k-1}|\mathcal{F}_{k-1}] = R^i_{k-1}$, where ΔM^i_k is the ith element of ΔM_k. Also, the R^i_k are bounded in L^2 because

$$\begin{aligned} E\left[(R^i_k)^2\right] &= E\left[\sum_{\ell=1}^{k}\rho(\ell)^{-1}(\Delta M^i_\ell)^2\right] \\ &= E\left[\sum_{\ell=1}^{k}\rho(\ell)^{-1}E\left[w^2_\ell(x^i_{\ell-1})^2|\mathcal{F}_{\ell-1}\right]\right] \\ &= \sum_{\ell=1}^{k}\rho(\ell)^{-1}E\left[(x^i_{\ell-1})^2\right] < \infty. \end{aligned}$$

Here $x^i_{\ell-1}$ is the ith element of $x_{\ell-1}$ and we have used that w_ℓ and $x_{\ell-1}$ are uncorrelated, that $E[w^2_\ell|\mathcal{F}_{\ell-1}] = 1$ and that for stable systems $(x^i_k)^2 < B_\infty$for all k, i for some $B_\infty < \infty$. By standard martingale results, [15, 13],

$$\lim_{k\to\infty} R^i_k = \lambda^i(\omega) < \infty \text{ a.s.} \qquad \text{for } i = 1,\ldots,N.$$

□

Lemma 2 *Suppose the system (1),(2) is stable then*

$$\lim_{k\to\infty}\rho(k)^{-\frac{1}{2}}\sum_{i=1}^{k}\Delta M_i = 0_{1\times N} \ a.s.$$

where $0_{1\times N}$ is the $(1 \times N)$ vector of all zeros. That is,

$$\lim_{k\to\infty}\rho(k)^{-\frac{1}{2}}M_k = 0_{1\times N} \ a.s.$$

Proof: Follows from Lemma 1 and by applying the Kronecker Lemma to each element of R_k [14, 13]. □

Theorem 2 *Consider the linear system (1),(2) and suppose the x_k and y_k are both observed. Suppose (16) and (17) hold. Then*

$$\lim_{k\to\infty}\hat{A}_k, \hat{B}_k, \hat{C}_k, \hat{D}_k = A, B, C, D \ a.s. \quad and$$

the almost sure convergence rate is at least $(k^{-2}\rho(k))^{\frac{1}{2}}$.

Proof: First consider the error in estimation of A, that is $\hat{A}_k - A = B(\frac{1}{k}M_k)(\frac{1}{k}O_k)^{-1}$. From Lemma 2 we have that

$$\lim_{k\to\infty} \frac{1}{k} M_k = 0_{1\times N} \text{ a.s.} \tag{18}$$

at a rate of $\frac{\rho(k)}{k^2}^{-\frac{1}{2}}$. Under the persistently exciting assumption $\lim_{k\to\infty}\sup_k \left(\frac{1}{k}O_k\right)^{-1}$ exists, so

$$\lim_{k\to\infty} \tilde{A}_k := \hat{A}_k - A = 0_{N\times N} \text{ a.s. .}$$

Indeed $\tilde{A}_k \to 0_{N\times N}$ a.s. at a rate $\frac{\rho(k)}{k^2}^{-\frac{1}{2}}$. The result for C follows similarly using that $x_{\ell-1}$ and $v_{\ell-1}$ are uncorrelated and that $E[v_k^2] = 1$.

The convergence results for B and D follow because $E[w_k^2] = 1$ or equivalently $\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} w_i^2 = 1$. □

Remarks

1. Similar convergence results for fully observed stochastic linear systems are common [5, 6].

Next we propose parameter estimators for partially-observed linear systems based on the conditional mean estimates of J_k, O_k and T_k. Convergence results are also presented.

3 Conditional Mean Estimates

Consider again the system (1), (2). We defined a model set Λ of allowable model estimates $\hat{\lambda}_k$ and assume that the true model $\lambda \in \Lambda$. The model set commonly considered in this paper is $\Lambda = \{\lambda(A, B, C, D, x_0) : N\text{th order model}\}$. Let us denote the associated conditional mean estimates based on the signal generating system λ as in (3), also termed the "correct" model, and the associated conditional mean estimates based on a possibly time varying estimate $\hat{\Lambda}_k = \{\hat{\lambda}_1, \ldots, \hat{\lambda}_k\}$, also termed the "incorrect" model, as, respectively,

$$\hat{J}_{k|k,\lambda} = E[J_k|\mathcal{Y}_k, \lambda], \quad \hat{O}_{k|k,\lambda} = E[O_k|\mathcal{Y}_k, \lambda], \quad \hat{T}_{k|k,\lambda} = E[T_k|\mathcal{Y}_k, \lambda] \tag{19}$$

and

$$\hat{J}_{k|k,\hat{\Lambda}_k} = E[J_k|\mathcal{Y}_k, \hat{\Lambda}_k], \quad \hat{O}_{k|k,\hat{\Lambda}_k} = E[O_k|\mathcal{Y}_k, \hat{\Lambda}_k], \quad \hat{T}_{k|k,\hat{\Lambda}_k} = E[T_k|\mathcal{Y}_k, \hat{\Lambda}_k] \tag{20}$$

In this section we consider parameter estimators, with the true model known or with the model estimates available, respectively. That is, we consider:

$$\hat{A}_{k|k,\lambda} = \hat{J}_{k|k,\lambda}\hat{O}^{-1}_{k|k,\lambda}, \quad \hat{C}_{k|k,\lambda} = \hat{T}_{k|k,\lambda}\hat{O}^{-1}_{k|k,\lambda} \tag{21}$$

$$\hat{B}_{k|k,\lambda} = E[B|\mathcal{Y}_k, \lambda], \quad \hat{D}_{k|k,\lambda} = E[D|\mathcal{Y}_k, \lambda] \tag{22}$$

and

$$\hat{A}_{k|k,\hat{\Lambda}_k} = \hat{J}_{k|k,\hat{\Lambda}_k}\hat{O}^{-1}_{k|k,\hat{\Lambda}_k}, \quad \hat{C}_{k|k,\hat{\Lambda}_k} = \hat{T}_{k|k,\hat{\Lambda}_k}\hat{O}^{-1}_{k|k,\hat{\Lambda}_k} \tag{23}$$

$$\hat{B}_{k|k,\hat{\Lambda}_k} = E[B|\mathcal{Y}_k, \hat{\Lambda}_k], \quad \hat{D}_{k|k,\hat{\Lambda}_k} = E[D|\mathcal{Y}_k, \hat{\Lambda}_k]. \tag{24}$$

Suppose the persistently excitation condition associated with the signal generating system λ and its estimate $\hat{\Lambda}_k$ holds, so that:

$$\lim_{k\to\infty}\sup_k \left(\frac{1}{k}\hat{O}_{k|k,\hat{\Lambda}_k}\right)^{-1} \quad \text{exists.} \tag{25}$$

Theorem 3 *Consider the linear system (1)(2), denoted by λ. Suppose (16) and (17) hold. Consider also an assumed that the conditional mean estimates based on the "correct model" are available which is persistently exciting, so that (25) holds. Then*

$$\lim_{k\to\infty} \hat{A}_{k|k,\lambda}, \hat{B}_{k|k,\lambda}, \hat{C}_{k|k,\lambda}, \hat{D}_{k|k,\lambda} = A, B, C, D \quad a.s. \quad and \tag{26}$$

the almost sure convergence rate is at least $(k^{-2}\rho(k))^{\frac{1}{2}}$.

Proof: We first proceed with the lemma result for $\hat{A}_{k|k,\lambda}$. Simple manipulations of (6), (20) and (23) give the error term, as:

$$\tilde{A}_{k|k,\lambda} = \hat{A}_{k|k,\lambda} - A \;=\; E[k^{-1}BM_k|\mathcal{Y}_k,\lambda]\left(E[k^{-1}O_k|\mathcal{Y}_k,\lambda]\right)^{-1} \tag{27}$$

From Lemma 2 write that $\phi_k := \rho(k)^{-\frac{1}{2}}M_k = \rho(k)^{-\frac{1}{2}}\sum_{i=1}^{k}\Delta M_i$, so that $\lim_{k\to\infty}\phi_k = \mathbf{0}_{1\times N}$ a.s.. Now, ϕ_k is bounded in L_2 since

$$\begin{aligned} E[\phi_k\phi_k'] &= E\left[\rho(k)^{-\frac{1}{2}}w_{k+1}x_k'x_kw_{k+1}\rho(k)^{-\frac{1}{2}}\right] \\ &= E\left[\rho(k)^{-1}E\left[w_{k+1}^2|\mathcal{F}_k\right]E\left[x_k'x_k|\mathcal{F}_k\right]\right] \\ &= \rho(k)^{-1}E\left[x_k'x_k\right] < \infty. \end{aligned}$$

Here we have used that for stable systems $x_k'x_k < B_\infty$ for all k for some $B_\infty < \infty$ so that $E[x_k'x_k|\mathcal{F}_k] < B_\infty$. Also $E\left[w_{k+1}^2|\mathcal{F}_k\right] = 1$.

This is a uniform integrability condition, which together with the property $\lim_{k\to\infty}\phi_k = \mathbf{0}_{1\times N}$ a.s. ensures convergence in conditional mean:

$$E[\phi_k|\mathcal{Y}_k,\lambda] \to \mathbf{0}_{1\times N} \quad a.s.$$

(Also $E[\phi_k] \to \mathbf{0}_{1\times N}$ a.s.). Hence, $E[k^{-1}M_k|\mathcal{Y}_k,\lambda] \to 0$ a.s. at a rate $(k^{-2}\rho(k))^{\frac{1}{2}}$. This gives the convergence (and rate) result (26) for $\hat{A}_{k|k,\lambda}$ under the excitation condition (25).

Similarly, the lemma convergence result holds for $\hat{C}_{k|k,\lambda}$.

Convergence for $\hat{B}_{k|k,\lambda}$ and $\hat{D}_{k|k,\lambda}$ follows because $E[w_k^2] = 1$. □

Remarks

1. In a companion paper we examine convergence for parameter estimators that use conditional mean estimates based on a sequence "incorrect" models that converges to the "correct" model using ordinary differential equation analysis.

2. The state space representation of a linear system is non-unique. Many representations are equivalent. In terms of the result of Theorem 3, the true system is one system in a class of equivalent models. The model actually estimated from a particular data set will be in the class of models equivalent to the "correct model". Which model is actually estimated from the class equivalent to the "correct model" will depend on the particular data set.

3. Optimal finite-dimensional filters for $\hat{O}_{k|k,\lambda}$ and $\hat{T}_{k|k,\lambda}$ are known and require $O(N^3)$ calculations per time instant[9]. We present a finite dimensional filter for $\hat{J}_{k|k,\lambda}$ in the next section. Sup-optimal filters can be implemented for $\hat{O}_{k|k,\lambda}$, $\hat{J}_{k|k,\lambda}$ and $\hat{T}_{k|k,\lambda}$ from Kalman filter state estimates. For example, $\hat{O}^{sub}_{k|k,\lambda} = \sum_{\ell=1}^{k}\hat{x}_{\ell-1}'\hat{x}_{\ell-1}$. Convergence results when sub-optimal estimates are used are neither included or excluded by convergence results of this paper.

3.1 Finite Dimensional Filter for J_k

In this section we derive a new finite dimensional filter for partially-observed stochastic linear systems. The techniques used here follow the derivation used in [9].

Let α_k and κ_k be the un-normalised densities defined as follows,

$$\alpha_k(x) := \bar{E}[\bar{\Gamma}_k I(x_k \in dx)|\mathcal{Y}_k], \quad \kappa_k(x)^{ij} := \bar{E}[\bar{\Gamma}_k J_k^{ij} I(x_k \in dx)|\mathcal{Y}_k],$$

The follow theorem gives recursive expressions for α_k and κ_k. The recursions are derived under the measure $\bar{P}$.

Theorem 4 *For $k \in \mathcal{Z}^+$, the un-normalised densities κ_k^{ij} and α_k are given by the following recursions:*

$$\begin{aligned} \kappa_k^{ij}(x) &= \frac{\phi(y_k - Cx)}{\phi(y_k)} \int_{R^N} \kappa_{k-1}^{ij}(x)\psi(x - Az)dz \\ &\quad + \frac{\phi(y_k - Cx)}{\phi(y_k)} \langle x, e_i \rangle \int_{R^N} \langle z, e_j \rangle \alpha_{k-1}(z)\psi(x - Az)dz && (28) \\ \alpha_k(x) &= \frac{\phi(y_k - Cx)}{\phi(y_k)} \int_{R} \alpha_{k-1}(z)\psi(x - Az)dz && (29) \end{aligned}$$

Proof: The recursion (29) is proven in [9]. We prove (28) here. Let $g : R^N \rightarrow R$ be any integrable Borel test function. So,

$$\begin{aligned} \bar{E}[\bar{\Gamma}_k g(x_k)|\mathcal{Y}_k] &= \int_{R^N} \alpha_k(x)g(x)dx \\ \bar{E}[\bar{\Gamma}_k J_k^{ij} g(x_k)|\mathcal{Y}_k] &= \int_{R^N} \kappa_k^{ij}(x)g(x)dx. && (30) \end{aligned}$$

Since $J_k^{ij} = J_{k-1}^{ij} + \langle x_k, e_i \rangle \langle x_{k-1}, e_j \rangle$, using the measure change we have

$$\begin{aligned} \bar{E}[\bar{\Gamma}_k J_k^{ij} g(x_k)|\mathcal{Y}_k] &= \bar{E}[\bar{\Gamma}_{k-1} \frac{\phi(y_k - Cx_k)}{\phi(y_k)} \frac{\psi(x_k - Ax_{k-1})}{\psi(x_k)} J_{k-1}^{ij} g(x_k)|\mathcal{Y}_k] \\ &\quad + \bar{E}[\bar{\Gamma}_{k-1} \frac{\phi(y_k - Cx_k)}{\phi(y_k)} \frac{\psi(x_k - Ax_{k-1})}{\psi(x_k)} \langle x_k, e_i \rangle \langle x_{k-1}, e_j \rangle g(x_k)|\mathcal{Y}_k] \\ &= \bar{E}[\bar{\Gamma}_{k-1} J_{k-1}^{ij} \left[\int_{R^N} \frac{\phi(y_k - Cx)}{\phi(y_k)} \psi(x - Ax_{k-1})g(x)dx \right] |\mathcal{Y}_{k-1}] \\ &\quad + \bar{E}[\bar{\Gamma}_{k-1} \langle x_{k-1}, e_j \rangle \left[\int_{R^N} \frac{\phi(y_k - Cx)}{\phi(y_k)} \psi(x - Ax_{k-1}) \langle x, e_i \rangle g(x)dx \right] |\mathcal{Y}_{k-1}] \\ &= \int_{R^N} \int_{R^N} \kappa_{k-1}^{ij} \frac{\phi(y_k - Cx)}{\phi(y_k)} \psi(x - Az)g(x)dxdz \\ &\quad + \int_{R^N} \frac{\phi(y_k - Cx)}{\phi(y_k)} \langle x, e_i \rangle g(x) \left[\int_{R^N} \langle z, e_j \rangle \psi(x - Az)\alpha_{k-1}(z)dz \right] dz && (31) \end{aligned}$$

where the second equality follows from independence of x_k and x_{k-1} under $\bar{P}$. Since $g(x)$ is an arbitrary Borel test function, equating (30) with (31) proves (28). □

Now we introduce the a finite dimension filter for J_k^{ij}. It was shown in [9] that densities for O_k and T_k could be characterised in terms of a finite number of sufficient statistics. The following theorem holds

Theorem 5 *At time k, the density $\kappa_k^{ij}(x)$ is complete defined by the 5 statistics a_k^{ij}, b_k^{ij}, d_k^{ij}, R_k and μ_k as follows:*

$$\kappa_k^{ij}(x) = [a_k^{ij} + b^{ij'}x + x'd_k^{ij}x]\alpha_k(x). \tag{32}$$

where $a_k^{ij} \in R$, $b_k^{ij} \in R^N$ and $d_k^{ij} \in R^{N\times N}$ with the pqth element denoted $d_k^{ij}(p,q)$ and the density $\alpha_k(x)$ is defined by the Kalman mean and variance statistics, μ_k and R_k. The statistics a_k^{ij}, b_k^{ij} and d_k^{ij} are given by the following recursions:

$$\begin{aligned}
a_{k+1}^{ij} &= a_k^{ij} + b_k^{ij'}\sigma_{k+1}^{-1}R_k^{-1}\mu_k + \sum_{p=1}^{N}\sum_{q=1}^{N} d_k^{ij}(p,q)\sigma_{k+1}^{-1}(p,q) \\
&\quad + \mu_k' R_k^{-1}\sigma_{k+1}^{-1} d_k^{ij}\sigma_{k+1}^{-1}R_k^{-1}\mu_k, \qquad a_0^{ij} = 0 & (33) \\
b_{k+1}^{ij} &= \Sigma_{k+1}^{-1}A\sigma_{k+1}^{-1}(b_k^{ij} + (d_k^{ij} + d_k^{ij'})\sigma_{k+1}^{-1}R_k^{-1}\mu_k) + e_i(e_j'\sigma_{k+1}^{-1}R_k^{-1}\mu_k), b_0^{ij} = 0_{N\times 1} & (34) \\
d_{k+1}^{ij} &= \Sigma_{k+1}^{-1}A\sigma_{k+1}^{-1}d_k^{ij}\sigma_{k+1}^{-1}A\Sigma_{k+1}^{-1} + \sum_{\ell=1}^{N} e_i(e_\ell'\Sigma_{k+1}^{-1}A\sigma_{k+1}^{-1}e_j)e_\ell', \qquad d_0^{ij} = e_ie_j' & (35)
\end{aligned}$$

where $\Sigma_k = B'B$ and $\sigma_k = A'\Sigma_k^{-1}A + R_{k-1}^{-1}$. The usual Kalman filter provides recursions for μ_k and R_k.

Proof: We prove (32) by induction. Assume at time k that $\kappa_k^{ij}(x)$ is of the form in (32). It follows at time $k+1$ using recursion (28) that $\kappa_{k+1}(x)$ is of the form of (32). To obtain recursions for a_k^{ij}, b_k^{ij} and d_k^{ij} we equate like terms. The tedious algebraic manipulations are left out here but the reader is referred to [9] for an example of working. □

Having characterised the density $\kappa_k^{ij}(x)$ by the finite sufficient statistics, we now derive a finite dimensional filter.

Theorem 6

$$E[J_k^{ij}|\mathcal{Y}_k] = a_k^{ij} + b_k^{ij'}\mu_k + \sum_{p=1}^{N}\sum_{q=1}^{N} d_k^{ij}(p,q)R_k(p,q) + \mu_k'd_k\mu_k \tag{36}$$

Proof: Using the abstract Bayes rule (5) we have

$$E[J_k^{ij}|\mathcal{Y}_k] = \frac{\bar{E}[\bar{\Gamma}_k J^{ij}|\mathcal{Y}_k]}{\bar{E}[\bar{\Gamma}_k|\mathcal{Y}_k]} = \frac{\int_{R^N}\kappa_k^{ij}(x)dx}{K} \tag{37}$$

where the constant $K = \int_{R^N}\alpha_k(x)dx$. It follows that

$$\begin{aligned}
\int_{R^N}\kappa_k^{ij}(x)dx &= KE[a_k^{ij} + b_k^{ij'}x + x'd_k^{ij}x|\mathcal{Y}_k] \\
&= K\left[a_k^{ij} + b_k^{ij'}\mu_k + \sum_{p=1}^{N}\sum_{q=1}^{N} d_k^{ij}(p,q)R_k(p,q) + \mu_k'd_k\mu_k\right] & (38)
\end{aligned}$$

substituting in (37) proves the theorem. □

Remarks:

1. This filter requires $O(N^5)$ calculations per time instant. Reducing computational effort is not investigated here.

2. Simpler recursions can be developed by noting that $E[x_{k-1}|\mathcal{Y}_{k-1}, x_k] = \int_{R^N}\langle z, e_j\rangle\psi(x - Az)\alpha_{k-1}(z)dz$ which can be approximated by $E[x_{k-1}|\mathcal{Y}_{k-1}]$.

4 Simulations

In this section, simulation studies are presented of a selection of the estimation schemes.

Ex 1: (Adaptive estimation) A 20000 point, 2-state linear system (1) (2) was generated with parametric values $A = \begin{bmatrix} 0 & -0.8 \\ 1 & -0.1 \end{bmatrix}$, $B = [1,0]'$, $C = [1,0.2]'$, $D = 0.01$. This system is in companion form and has eigenvalues of $-0.0500 \pm 0.8930i$. In this simulation we assumed that C and D are known and A and B are estimated by (20) and (23) Our random initial guess for the model is, $\hat{A}_0 = \left[\begin{array}{cccc} 0.9347 & 0.5194; & 0.3835 & 0.8310 \end{array}\right]$ which has eigenvalues of 1.3321 and 0.4335. The model estimate $\hat{\lambda}_k$, used in the J_k, O_k filters, is updated at each time instant by the new parameter estimates $\hat{A}_{k|k,\hat{\lambda}_k}$. Note that the estimated system has also been transformed into companion form at each time step. Figure 1 shows convergence of $\hat{a}_{21}$. The parameter estimates for the other elements of A and other system parameters converge similarly. The final estimated system has eigenvalues of $-0.0361 \pm 0.8810i$, which compares well with the generating system. Note that the parameter estimators are not turned on until after 1000 points to allow J_k and O_k filters to forget initial conditions.

Ex 2: (More Noise) The same process is generated, except that $D = 1$ (more measurement noise). In this simulation we assumed that C and D are known and A and B are estimated by (20) and (23) Our initial guess is $\hat{A}_0 = \left[\begin{array}{cccc} 0.0346 & 0.5297; & 0.0535 & 0.6711 \end{array}\right]$ which has eigenvalues of 1.3321 and 0.4335. The model estimate $\hat{\lambda}_k$, used in the J_k, O_k filters, is updated at each time instant by new parameter estimates $\hat{A}_{k|k,\hat{\lambda}_k}$. The estimated value of A after 20000 points is $\hat{A} = \left[\begin{array}{cccc} -0.0060 & -0.7949; & 0.9772 & -0.0810 \end{array}\right]$ Figure 2 shows convergence of $\hat{a}_{12}$. The system is transformed into companion form at each time step.

5 Conclusions

Global convergence results have been developed for new finite dimensional adaptive schemes for estimation of the parameters of a partially observed discrete-time linear stochastic system. Preliminary convergence results are developed in this paper using standard martingale properties and convergence results. We emphasize that, for stable linear systems driven by white noise, there is consistent estimation and this is asymptotically optimal. Almost-sure rates of convergence also are guaranteed to be of the order $(k^{-2}\rho(k))^{\frac{1}{2}}$, where $\rho(k) = max(1, klnk(lnlnk)^{\alpha})$, for any $\alpha > 1$. More complete results are developed in a companion paper.

References

[1] R.J. Elliott, L. Aggoun and J.B. Moore, *Hidden Markov Models, Estimation and Control*, Springer-Verlag, 1995.

[2] P.R. Kumar and P Varaiya, *Stochastic Systems: Estimation, Identification, and adaptive control*, Prentice-Hall, Englewood Cliffs, New Jersey, 1986.

[3] L. Ljung, *Identification: Theory for the user*, Prentice Hall, Englewood Cliffs, New Jersey, 1987.

[4] L. Ljung, *Theory and Practice of Recursive identification*, MIT press, England, 1983.

[5] N. Kalouptsidis and S. Theodoridis, *Adaptive System Identification and Signal Processing Algorithms*, Prentice Hall, New York, 1993.

[6] J.S. Lim and A.V. Oppenheim, *Advanced Topics in Signal Processing*, Prentice Hall, Englewood Cliffs, 1998.

[7] S.T. Alexander, *Adaptive Signal Processing: Theory and Applications*, Springer-Verlag, New York, 1986.

[8] T. Soderstrom and P. Stocia, *System Identification*, Prentice Hall, New York, 1989.

[9] R.J. Elliott and V. Krishnamurthy, "New Finite Dimensional Filters for Estimation of Linear Gauss-Markov Models", to appear in *Proc. of CDC97.*

[10] R.J. Elliott, J. Ford and J.B. Moore, "On-line Consistent Estimation of Hidden Markov Models", submitted to *IEEE Trans. on Information Theory.*

[11] R.J. Elliott and J.B. Moore, "Almost Sure Parameter Estimation and Convergence rates for Hidden Markov Models", to appear in *Systems and Control Letters.*

[12] R.J. Elliott and J.B. Moore, "A Martingale Kronecker Lemma and Parameter Estimation for Linear Systems", to appear in *IEEE Trans. in Automatic Control.*

[13] J. Neveu, *Discrete Parameter Martingales*, Amsterdam, Holland: North-Holland, 1975.

[14] M. Loeve, *Probability Theory*, 2nd Ed., Van Nostrand, Princeton, NJ, 1960.

[15] P. Meyer, *Martingales and Stochastic Intergrals -I*, Lecture Notes in Mathematics Series, No 284, Springer-Vertag, New York, 1972.

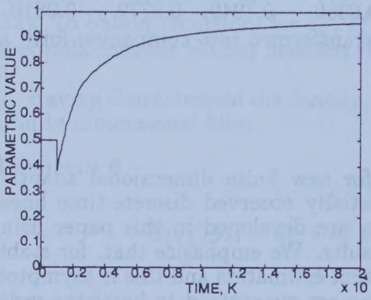


Figure 1: Adaptive Estimation

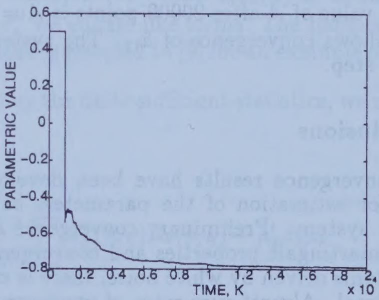


Figure 2: More noise example

Weight reoptimization in OBE algorithms: An initial study

D. Joachim, J.R. Deller, Jr.[1]*, and M. Nayeri*
Michigan State University
Department of Electrical Engineering / 2120 EB
Speech Processing and Adaptive Signal Processing Laboratories
East Lansing, MI 48824-1226 USA

Abstract. An enhanced *optimal bounding ellipsoid* algorithm is presented in which multiple weights are optimized when data are selected for updating. Minimizing the OBE cost functions over multiple weights potentially offers enhanced performance in terms of convergence rate, tracking ability, and solution quality. The enhanced algorithm is investigated with respect to these issues through simulation studies.

1 OBE with weight reoptimization

At the heart of *optimal bounding ellipsoid* (OBE) identification algorithms is an iterative weight assignment to each incoming data vector that reflects the current observation's potential to refine the solution set [1]. Each weight is determined by minimizing a measure of the size of an hyperellipsoidal feasibility set to which the "true" parameter vector must belong. This optimization takes place pointwise; all conditions (in particular, previous weights) extant at the time of optimization remain fixed. The globally optimal solution at a given time would diminish the solution set in light of *all* observations taken until that time. In this work, we develop an OBE algorithm that "revisits" K past weights when data at time n are received, so that the ellipsoid is optimally diminished with respect to the current and past K observations, conditioned upon information known at time $n-K-1$.

OBE algorithms are used to identify *linear-in-parameters* models of the form

$$y_n = \boldsymbol{\theta}_*^T \boldsymbol{x}_n + \varepsilon_{n*} \tag{1}$$

in which $\boldsymbol{\theta}_* \in \Re^m$ is the unknown "true" parameter vector to be identified; $\{\boldsymbol{x}_n\}$ is a sequence of measurable vectors of dimension m; and $\{\varepsilon_{n*}\}$ is an error sequence. OBE algorithms are based on the premise that, for each n, the model error has a known pointwise energy bound,

$$\varepsilon_{n*}^2 \leq \gamma_n^2. \tag{2}$$

Given data on times $t \in [1, n]$, therefore, an *exact feasibility set*, say Ω_n, of estimates for $\boldsymbol{\theta}_*$ whose elements are consistent with these bounds is formally described by

$$\Omega_n = \cap_{t=1}^n \omega_t, \quad \text{where} \quad \omega_t = \{\boldsymbol{\theta} : \varepsilon_{t*}^2 = |y_t - \boldsymbol{\theta}^T \boldsymbol{x}_t|^2 \leq \gamma_t^2\}. \tag{3}$$

OBE algorithms work with an hyperellipsoidal set, say $\bar{\Omega}_n$, that is guaranteed to contain Ω_n, hence $\boldsymbol{\theta}_*$. The observations are scrutinized with respect to their ability to "shrink" $\bar{\Omega}_n$, hence to tightly bound Ω_n. At time n, the hyperellipsoid is given by (e.g., [1])

$$\bar{\Omega}_n \overset{\text{def}}{=} \left\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\theta}_n)^T \boldsymbol{C}_n (\boldsymbol{\theta} - \boldsymbol{\theta}_n) \leq \kappa_n\right\} \tag{4}$$

---

[1] This work was completed in part while JD was on leave at the University of Notre Dame. He is indebted to Profs. Y.F. Huang and D.J. Costello, and to doctoral candidates S. Golamudi, S. Nagaraj, and S. Kapoor for many interesting and informative discussions about the QOBE algorithm which serves as the basis for the new OBE algorithm developed here.

in which $\boldsymbol{C}_n$ is the weighted *covariance* matrix of the observations,

$$\boldsymbol{C}_n = \sum_{t=1}^{n} q_{t,n}\boldsymbol{x}_t\boldsymbol{x}_t^T, \tag{5}$$

κ_n is the scalar

$$\kappa_n = \boldsymbol{\theta}_n^T\boldsymbol{C}_n\boldsymbol{\theta}_n + \sum_{t=1}^{n} q_{t,n}(\gamma_t^2 - y_t^2), \tag{6}$$

and $\boldsymbol{\theta}_n$, the center of $\bar{\Omega}_n$, is the weighted least-square-error estimator of $\boldsymbol{\theta}_*$ at time n,

$$\boldsymbol{\theta}_n = \boldsymbol{P}_n\boldsymbol{c}_n, \quad \text{with } \boldsymbol{P}_n \overset{\text{def}}{=} \boldsymbol{C}_n^{-1} \text{ and } \boldsymbol{c}_n \overset{\text{def}}{=} \sum_{t=1}^{n} q_{t,n}y_t\boldsymbol{x}_t. \tag{7}$$

The weighting sequence in this process at time n, $\{q_{t,n}\}_{t=1}^n$, is chosen to optimally diminish some set measure of the hyperellipsoid (see Section 3).

OBE algorithms make selective use of incoming data in updating the ellipsoid and central estimator. Frequently, the observations at time n contain no innovation in the sense that they cannot be used to reduce the size of $\bar{\Omega}_n$. This is manifest in the failure to find valid weights, and the effort of updating can be avoided at this time. Depending on the properties of the sequence $\{\varepsilon_{n*}\}$, OBE algorithms often update only 10 percent of the time or less.

All published OBE algorithms can be manipulated into the formal framework described above provided that we allow for time-varying ("n-dependent") weight sequences as we have done in (5) - (7) [1]. In all published cases, however, this time dependence (if any) has a simple structure arising from a either a generalized "forgetting factor" (e.g., [2, 3]) or some heuristic measures to induce adaptation (e.g., [4]). In no published case is there an attempt to *reoptimize* any of the "previous" weights at time n in light of the new measurements $\boldsymbol{x}_n$ and y_n. That is, all optimization in existing OBE algorithms can be accomplished by manipulation the *current* weight only. This can be inferred directly from the work in [1]. The globally optimal solution at time n would optimize *all* weights $\{q_{t,n}\}_{t=1}^n$, in light of *all* known information $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^n$.

In this work, we develop an algorithm that can potentially approach the performance of a globally-optimized algorithm, by "revisiting" $K > 0$ past weights when data at time n are received. The revision of past weights is made by additive adjustments to existing weight values, subject to the constraint that the revised weight remain nonnegative. Accordingly, the time-varying weights are computed as

$$q_{t,n} = \begin{cases} \sum_{i=0}^{K} \lambda_t^i, & 0 \leq t \leq n-K \\ \sum_{i=0}^{n-t} \lambda_t^i, & n-K < t \leq n \end{cases}, \tag{8}$$

with the constraint $q_{t,n} \geq 0$ for any t and n. λ_t^i for $i > 0$ denotes the ith additive adjustment to the weight at time t, and λ_n^0 is the original weight computation at time n. The resulting algorithm, therefore, optimally diminishes the hyperellipsoid size (in a sense to be described) with respect to the current and past K observations, conditioned upon quantities computed through time $n - K - 1$.

The reoptimized OBE algorithm characteristically checks to assure that the incoming data are informative in their potential to diminish the set $\bar{\Omega}_n$. This effectively checks for the existence of an optimal $\lambda_n^0 > 0$ (see Section 4). If the new observation is not innovative, then the past weights are already optimal in light of the new measurement

Table 1: Summary of vector and matrix notation.

- $A(i,j)$ denotes the (i,j) element of matrix $\boldsymbol{A}$ and $a(i)$ the ith element of vector $\boldsymbol{a}$.
- If $\boldsymbol{a}$ is an N-vector, $\boldsymbol{A} = \text{diag}(\boldsymbol{a})$ is the diagonal matrix with ith diagonal element $a(i)$.
- If $\boldsymbol{A}$ is an $N \times M$ matrix, $\boldsymbol{A}(*, i)$ is the N-vector comprising the ith column of $\boldsymbol{A}$.
- The *Hadamard product* of two $N \times M$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted $\boldsymbol{A} \circ \boldsymbol{B}$, is defined to be the $N \times M$ matrix with (i,j) element $A(i,j)B(i,j)$ [5].

Quantity	Notation	Dimensions
Regressor matrix	$\boldsymbol{X}_n \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{x}_{n-K} & \cdots & \boldsymbol{x}_{n-1} & \boldsymbol{x}_n \end{bmatrix}$	$m \times K$
Output vector	$\boldsymbol{y}_n \overset{\text{def}}{=} \begin{bmatrix} y_{n-K} & \cdots & y_{n-1} & y_n \end{bmatrix}^T$	$K \times 1$
Weight vector	$\boldsymbol{\lambda}_n \overset{\text{def}}{=} \begin{bmatrix} \lambda^{k-1}_{n-K+1} & \cdots & \lambda^1_{n-1} & \lambda^0_n \end{bmatrix}^T$	$K \times 1$
Weight matrix	$\boldsymbol{\Lambda}_n \overset{\text{def}}{=} \text{diag}(\boldsymbol{\lambda}_n)$	$K \times K$
Bound vector	$\boldsymbol{\gamma}_n \overset{\text{def}}{=} \begin{bmatrix} \gamma_{n-K} & \cdots & \gamma_{n-1} & \gamma_n \end{bmatrix}^T$	$K \times 1$
Bound matrix	$\boldsymbol{\Gamma}_n \overset{\text{def}}{=} \text{diag}(\boldsymbol{\gamma}_n)$	$K \times K$
Cond'l pred'n error vector	$\boldsymbol{\varepsilon}_n \overset{\text{def}}{=} \begin{bmatrix} \varepsilon_{n-K\|n-1} & \cdots & \varepsilon_{n\|n-1} \end{bmatrix}^T$	$K \times 1$
Instantaneous covariance-weighted regressor energy matrix	$\boldsymbol{G}_n \overset{\text{def}}{=} \boldsymbol{X}_n^T \boldsymbol{P}_{n-1} \boldsymbol{X}_n$	$K \times K$

and no updating is required. If the observation at time n is informative, then, once the set $\{\lambda^i_{n-i}\}_{i=0}^K$ is found, the update equations for $\boldsymbol{\theta}_n$, $\boldsymbol{C}_n$, and κ_n are relatively straightforward generalizations of the recursions used for "conventional" OBE. We now briefly discuss the development of these recursions before considering the more challenging optimization problem.

2 Reoptimization recursions

The vector and matrix notation defined in Table 1 is employed throughout the remaining developments.

An update in the reoptimized OBE algorithm can be viewed as a block extension of the conventional OBE updating process in which $K+1$ "new" (the current plus the reconsideration of K previous) observations are incorporated at the update time. Since two of the three key OBE recursions are *structurally* very similar to the well-known RLS algorithm [6], it is not surprising that their counterparts in the reoptimized OBE algorithm are similar to block-updated RLS.[2]

The update recursion for the *inverse* covariance matrix, $\boldsymbol{P}_n$, is obtained by applying the matrix-inversion lemma [8] to the readily-derived recursion for the covariance update $\boldsymbol{C}_n = \boldsymbol{C}_{n-1} + \boldsymbol{X}_n \boldsymbol{\Lambda}_n \boldsymbol{X}_n^T$, yielding

$$\boldsymbol{P}_n = \boldsymbol{P}_{n-1} - \boldsymbol{P}_{n-1} \boldsymbol{X}_n \left[\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{G}_n\right]^{-1} \boldsymbol{X}_n^T \boldsymbol{P}_{n-1}. \tag{9}$$

By similarly familiar mechanics, we derive the accompanying update recursions for $\boldsymbol{\theta}_n$

---

[2] An alternative is to employ the "QR decomposition" OBE recursions suggested by Deller *et al.* [4, 7].

and κ_n,

$$\begin{aligned} \boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} + \boldsymbol{P}_n \boldsymbol{X}_n \boldsymbol{\Lambda}_n \boldsymbol{\varepsilon}_n & (10)\\ \kappa_n &= \kappa_{n-1} + \boldsymbol{\gamma}_n^T \boldsymbol{\Lambda}_n \boldsymbol{\gamma}_n - \boldsymbol{\varepsilon}_n^T \boldsymbol{H}_n^{-1} \boldsymbol{\Lambda}_n \boldsymbol{\varepsilon}_n\,, \quad \boldsymbol{H}_n \overset{\text{def}}{=} \boldsymbol{I} + \boldsymbol{\Lambda}_n \boldsymbol{G}_n\,. & (11) \end{aligned}$$

3 Optimization

The matrix defining the hyperellipsoid at time n is given by [see (4)] $\boldsymbol{P}_n^{-1}/\kappa_n$. The measure det $\{\kappa_n \boldsymbol{P}_n\}$ is proportional to the square of the volume of the ellipsoid, and is most often minimized in the OBE optimization process. Minimization of the trace of $\{\kappa_n \boldsymbol{P}_n\}$ is also a meaningful measure of size (e.g., [1, 9]). Minimization of the parameter κ_n, first suggested in [2], has been controversial with respect to its interpretability [1] until recently when Huang's research group developed the *quasi-OBE* (QOBE) algorithm [10, 11]. QOBE, which minimizes κ_n in conjunction with a specific weighting strategy, provides interesting interpretations of this optimization process. Because of the relative algebraic simplicity of the algorithm, we use the QOBE-like approach of κ_n minimization in developing an initial reoptimized OBE alg

In the present work, therefore, the objective is to reduce κ_n at each iteration with respect to the diagonal elements of $\boldsymbol{\Lambda}_n$, if the new data admit further reduction. We begin with the recursion for κ_n in (11). Differentiating the term $\boldsymbol{H}_n^{-1} \boldsymbol{\Lambda}_n$ with respect to a nominal weight λ_{n-i}^i, we obtain

$$\frac{\partial(\boldsymbol{H}_n^{-1} \boldsymbol{\Lambda}_n)}{\partial \lambda_{n-i}^i} = \frac{\partial \boldsymbol{H}_n^{-1}}{\partial \lambda_{n-i}^i} \boldsymbol{\Lambda}_n + \boldsymbol{H}_n^{-1} \frac{\partial \boldsymbol{\Lambda}_n}{\partial \lambda_{n-i}^i} = \boldsymbol{H}_n^{-1} \frac{\partial \boldsymbol{\Lambda}_n}{\partial \lambda_{n-i}^i} \left(\boldsymbol{H}_n^{-1}\right)^T \tag{12}$$

from which we conclude that

$$\boldsymbol{\varepsilon}_n^T \frac{\partial \left(\boldsymbol{H}_n^{-1} \boldsymbol{\Lambda}_n\right)}{\partial \lambda_{n-i}^i} \boldsymbol{\varepsilon}_n = \boldsymbol{\varepsilon}_n^T \boldsymbol{H}_n^{-1}(*, i) \left[\boldsymbol{H}_n^{-1}(*, i)\right]^T \boldsymbol{\varepsilon}_n \tag{13}$$

where $\boldsymbol{H}_n^{-1}(*, i)$ is the ith column of matrix $\boldsymbol{H}_n^{-1}$. Equation (13) leads to

$$\begin{aligned} \frac{\partial \kappa_n}{\partial \lambda_{n-i}^i} &= {\gamma_{n-i+1}}^2 - \boldsymbol{\varepsilon}_n^T \boldsymbol{H}_n^{-1}(*, i) \left[\boldsymbol{H}_n^{-1}(*, i)\right]^T \boldsymbol{\varepsilon}_n \\ &= \left(\gamma_{n-i+1} + \boldsymbol{\varepsilon}_n^T \boldsymbol{H}_n^{-1}(*, i)\right)\left(\gamma_{n-i+1} - \boldsymbol{\varepsilon}_n^T \boldsymbol{H}_n^{-1}(*, i)\right). & (14) \end{aligned}$$

Hence,

$$\frac{\partial \kappa_n}{\partial \boldsymbol{\Lambda}_n} = \boldsymbol{\Gamma}_n - \left[\operatorname{diag}(\boldsymbol{H}_n^{-T} \boldsymbol{\varepsilon}_n)\right]^2 \tag{15}$$

where $\frac{\partial \kappa_n}{\partial \boldsymbol{\Lambda}_n}$ is the diagonal matrix whose ith diagonal element is $\frac{\partial \kappa_n}{\partial \lambda_{n-i}^i}$.

In terms of the Hadamard product notation (see Table 1) and (15), the optimal weight matrix is the solution of the equation

$$\boldsymbol{0} = \left(\boldsymbol{\gamma}_n + \boldsymbol{H}_n^{-T} \boldsymbol{\varepsilon}_n\right) \circ \left(\boldsymbol{\gamma}_n - \boldsymbol{H}_n^{-T} \boldsymbol{\varepsilon}_n\right) \tag{16}$$

which, using $|\boldsymbol{A}|$ to denote matrix with (i, j) element $|A(i, j)|$, becomes

$$|\boldsymbol{H}_n^{-T} \boldsymbol{\varepsilon}_n| - \boldsymbol{\gamma}_n = \boldsymbol{0}\,. \tag{17}$$

Let us define the column vector $\boldsymbol{s}_n \overset{\text{def}}{=} \operatorname{sign}\{\boldsymbol{H}_n^{-T}\boldsymbol{\varepsilon}_n\}$, and denote by $\boldsymbol{g}_n$ the column vector whose elements are the diagonal elements of $\boldsymbol{\Gamma}_n^{-1}$. Then

$$|\boldsymbol{H}_n^{-T}\boldsymbol{\varepsilon}_n| \;=\; (\boldsymbol{H}_n^{-T}\boldsymbol{\varepsilon}_n)\circ\boldsymbol{s}_n, \tag{18}$$

from which an the optimal weight vector is found to be the solution in $\boldsymbol{\lambda}_n$ of

$$\begin{aligned}\boldsymbol{\varepsilon}_n &= \boldsymbol{H}_n^T(\boldsymbol{\gamma}_n\circ\boldsymbol{s}_n)\\ &= (\boldsymbol{\gamma}_n\circ\boldsymbol{s}_n)+\boldsymbol{G}_n\boldsymbol{\Lambda}_n(\boldsymbol{\gamma}_n\circ\boldsymbol{s}_n).\end{aligned} \tag{19}$$

After some manipulation, we obtain

$$\boldsymbol{\lambda}_n \;=\; \left[\boldsymbol{G}_n^{-1}(\boldsymbol{\varepsilon}_n-\boldsymbol{\gamma}_n\circ\boldsymbol{s}_n)\right]\circ(\boldsymbol{g}_n\circ\boldsymbol{s}_n). \tag{20}$$

Further, if we define the matrix $\boldsymbol{S}_n \overset{\text{def}}{=} \operatorname{diag}(\boldsymbol{s}_n)$, then (20) can be also expressed as

$$\boldsymbol{\lambda}_n \;=\; (\boldsymbol{G}_n\boldsymbol{S}_n\boldsymbol{\Gamma}_n)^{-1}(\boldsymbol{\varepsilon}_n-\boldsymbol{S}_n\boldsymbol{\gamma}_n) \tag{21}$$

which is the desired solution for the weight vector. It is noteworthy that when $K=0$, this expression reduces to that the for the optimal scalar weight in the QOBE algorithm [10, 11].

4 Algorithm

The resulting algorithm appears in Table 2. The initial conditions mirror those in QOBE supplemented by the initial "accumulated" weight vector $\boldsymbol{\Sigma}_n$ which is appropriately assumed to contain all zero elements. This vector contains the accumulation of all adjustments made to the weights over the window of time that it currently represents, $n-K,\ldots,n$. At each iteration, the elements of $\boldsymbol{\Sigma}_n$ are "shifted" to reflect the new time window.

The check for innovation at time n in the conventional QOBE algorithm ($K=0$) is

$$|\varepsilon_n| > \gamma_n\,. \tag{22}$$

With weight reoptimization ($K>0$), satisfaction of this simple test is still necessary for any further computation to be required on the window $n-K,\ldots,n$. Indeed, if there is no innovation in the observation at time n, then the past K weights are already optimal. The computational cost of the method is therefore drastically reduced by recourse to the simple check (22) prior to further optimization. The computational complexity of the reoptimized algorithm is significantly worse *per update* than that of QOBE without reconsideration of past data. However, these selective algorithms tend to incorporate so few data that, even with the additional burden at times of update, the overall complexity remains $\mathcal{O}(m)$, the complexity of the computations at a time for which no update occurs. Further, there is empirical evidence that reoptimization may result in a significant reduction in the number of updates (Section 5). Finally, further work will almost certainly make the reoptimization process more efficient, since we have made no attempt to exploit properties such as matrix symmetries in the present algorithm.

5 Simulation Study

In the following simulation, we investigate the performance of the reoptimized OBE algorithm with a particular interest in the effects of K, the numbers of previous weight

Table 2: The QOBE algorithm with weight reoptimization.

```
I. Initialization:
    1. $\boldsymbol{\theta}_K = \mathbf{0}$, $\kappa_K = \mu$ and $\boldsymbol{P}_K = \frac{1}{\mu^2}\boldsymbol{I}$, $\mu$ small.
    2. $\boldsymbol{\Sigma}_K = \mathbf{0}$
II. Recursion:
    For $n = K+1, K+2, \ldots$
        $\Sigma_n = \left[\; \Sigma_{n-k,k-1} \;\; \Sigma_{n-k+1,k-2} \;\; \ldots \;\; \Sigma_{n-1,0} \;\; 0 \;\right]$
        $\varepsilon_{n|n-1} = y_n - \boldsymbol{\theta}_{n-1}^T \boldsymbol{x}_n$
        If $|\varepsilon_{n|n-1}| > \gamma_n$
            $\boldsymbol{\varepsilon}_n = \boldsymbol{y}_n - \boldsymbol{X}_n^T \boldsymbol{\theta}_{n-1}^T$
            If $|\boldsymbol{\varepsilon}_n| > \boldsymbol{\gamma}_n$
                $\boldsymbol{G}_n = \boldsymbol{X}_n^T \boldsymbol{P}_{n-1} \boldsymbol{X}_n$
                $\boldsymbol{\lambda}_n = \boldsymbol{\Gamma}_n^{-1} \boldsymbol{G}_n^{-1} (\boldsymbol{\varepsilon}_n - \boldsymbol{\gamma}_n \circ \mathrm{s}_n)$
                If $\boldsymbol{\lambda}_n + \Sigma_n \geq 0$
                    $\boldsymbol{P}_n = \boldsymbol{P}_{n-1} - \boldsymbol{P}_{n-1} \boldsymbol{X}_n \boldsymbol{H}_n^{-1} \boldsymbol{\Lambda}_n \boldsymbol{P}_{n-1}$
                    $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \boldsymbol{P}_n \boldsymbol{X}_n \boldsymbol{\Lambda}_n \boldsymbol{\varepsilon}_n$
                    $\kappa_n = \kappa_{n-1} + \boldsymbol{\gamma}_n^T \boldsymbol{\Lambda}_n (\boldsymbol{\gamma}_n - s_n \circ \boldsymbol{\varepsilon}_n^T)$
```

reoptimizations. We seek to identify an order-three autoregressive model following the difference equation

$$y(n) = -1.23y(n-1) - 0.64y(n-2) + 0.14y(n-3) + \varepsilon_{n*}, \tag{23}$$

where $\{\varepsilon_{n*}\}$ is biased and bounded by $[-0.5, +1]$.

Figure 1 shows the convergence behavior of the parameter estimator with values $K = 0, 1$ and 2. Recall that $K = 0$ corresponds to "conventional" QOBE. We observe that the convergence rate is slightly decreased with increased reoptimization of past weights. This is due to fact that the aggregate (i.e., over all $K+1$ data) checking criterion, which is quite conservative in the present version of the algorithm, is less often satisfied as K increases. We speculate that another factor might be at play as well. Nagaraj *et al.* [11] have found a tendency of κ_n minimization to result in a quick determination of a good central estimator. This tendency might be enhanced with weight reoptimization. In any case, the consequence of this behavior is drastically decreased updating (63 updates in 300 points when $K = 0$, 9 updates when $K = 2$), a phenomenon that offsets some of the extra computational burden of reoptimization. Decreased frequency of updating also has other important benefits, in freeing up the processor to share cycles with other tasks for instance [4, 10].

The relationship between κ_n minimization and the volume of the hyperellipsoid is not well-defined. It can be inferred from recent work, however, that κ_n minimization will reduce volume at each update [1, 10]. Figure 2 illustrates potential for significant improvement in hyperellipsoid volume reduction with increasing K. This effect would be expected to be even more pronounced in a reoptimized OBE algorithm based on volume minimization. Such an algorithm will be the subject of future research.

Finally, Figure 3 illustrates that κ_n is significantly decreased with increased K. This is entirely expected because more effective κ_n optimization is the numerical objective of the reoptimization process.

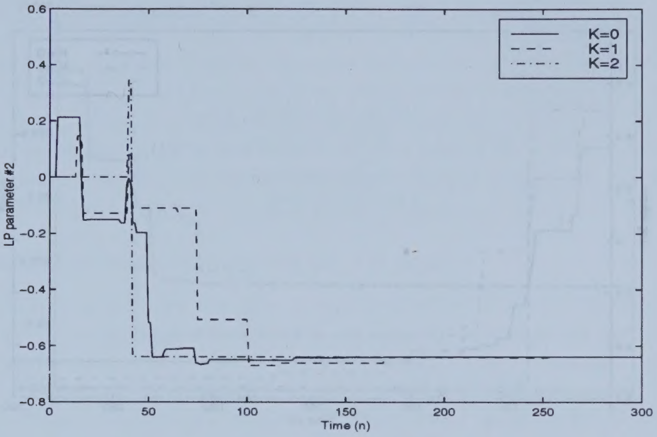


Figure 1: Parameter #2 plots comparing the performance for cases $K = 0, 1$ and 2 for the simulation study of Section 5.

The first 25 weight assignments in the simulation described above are reported in Table 3. The decreasing frequency of updates with increased K is clearly evident in the results.

6 Conclusion

This paper has presented an OBE algorithm that potentially reoptimizes K past weight assignments at each time. This initial algorithm has been based on "κ_n-minimization" and reduces to the QOBE algorithm [11] when $K = 0$. Simulation studies have revealed expected enhanced ability to minimize κ_n with increasing K. Enhanced ability to reduce volume also results with increasing K, although the precise mechanism for this phenomenon is not clear when κ_n is minimized. Reoptimization results in an increased computational cost per update, but the overall complexity remains $\mathcal{O}(m)$ since updating is very infrequent. In fact, the number of updates decreased dramatically with increasing K in the present simulations, but this phenomenon might be due in, at least in part, to a conservative aggregate check for innovation. Future work includes decreasing the algorithmic complexity and a study of the reoptimized performance in the volume-minimizing OBE algorithms.

References

[1] J. R. Deller, Jr., M. Nayeri, and M. Liu, "Unifying the landmark developments in optimal bounding ellipsoid identification," *Int. Adaptive Control and Signal Processing*, vol. 8, pp. 43–60, 1994.

[2] S. Dasgupta and Y.-F. Huang, "Asymptotically convergent modified recursive least squares with data dependent updating and forgetting factor for systems with bounded noise," *IEEE Trans. on Information Theory*, vol. IT-33, pp. 383–392, 1987.

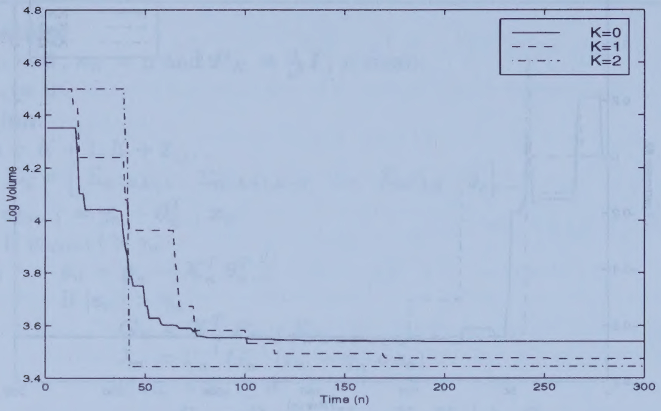


Figure 2: Volume plots comparing the performance for cases $K = 0, 1$ and 2 for the simulation study of Section 5.

[3] M. Nayeri, M. Liu, and J. R. Deller, Jr., "Do interpretable optimal bounding ellipsoid algorithms converge? parts I & II," in *Proc. 10th IFAC/IFORS Symp. on System Identification (SYSID '94)*, vol. 3, (Copenhagen), pp. 389–400, July 1994.

[4] J. R. Deller, Jr. and S. F. Odeh, "Adaptive set-membership identification in $\mathcal{O}(\mathrm{m})$ time for linear-in-parameters models," *IEEE Trans. on Signal Processing*, vol. 41, May 1993.

[5] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, vol. 2. Cambridge University Press, 1991.

[6] S. Haykin, *Adaptive Filter Theory*. Englewood-Cliffs, New Jersey: Prentice-Hall, 1995.

[7] J. R. Deller, Jr. and T. C. Luk, "Linear prediction analysis of speech based on set-membership theory," *Computer Speech and Language*, vol. 3, pp. 301–327, Oct. 1989.

[8] B. Noble, *Applied Linear Algebra*. Englewood-Cliffs, New Jersey: Prentice-Hall, 1969.

[9] E. Fogel and Y.-F. Huang, "On the value of information in system identification - Bounded noise case," *Automatica*, vol. 18, pp. 229–238, 1982.

[10] S. Gollamudi, S. Nagaraj, and Y.-F. Huang, "SMART: A toolbox for set-membership filtering," in *Proc. 1997 European Conf. on Circuit Theory and Design*, (Budapest), Sept. 1997. In press.

[11] S. Nagaraj, S. Gollamundi, J. R. Deller, Jr., S. Kapoor, and Y.-F. Huang, "Convergence analysis of a new 'Quasi-OBE' algorithm for real-time signal processing," in *Proc. 40th Annual Midwest Symposium on Circuits and Systems*, (Sacramento), Aug. 1997. In press.

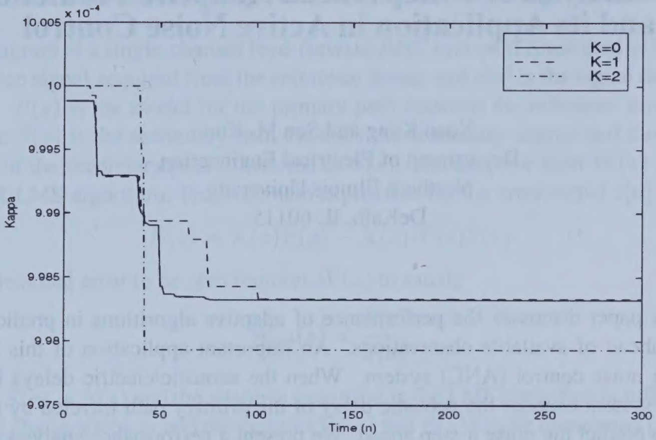


Figure 3: κ_n behavior parameterized by K for the simulation study of Section 5.

Table 3: First 25 weight assignments for cases $K = 0, 1$ and 2 in the simulation study of Section 5. The number of updates is 63, 24, 9 for K = 0, 1, 2 respectively.

Time,n	$\lambda_n^0, (K=0)$	$\lambda_n^0, (K=1)$	$\lambda_n^1, (K=1)$	$\lambda_n^0, (K=2)$	$\lambda_n^1, (K=2)$	$\lambda_n^2, (K=2)$
5	-	-	1.000e+00	-	1.000e+00	-
6	-	-	-	-	-	1.000e+00
7	-	-	-	-	-	1.000e+00
13	-	2.091e-07	-	-	-	-
15	-	-	2.091e-07	-	-	-
16	2.695e-07	4.469e-07	-	-	-	-
18	-	-	4.469e-07	-	-	-
38	-	-	-	3.249e-07	-	-
39	5.289e-07	2.340e-07	-	-	-	-
40	1.623e-07	-	-	2.428e-07	3.249e-07	1.701e-07
41	6.422e-08	1.324e-07	2.340e-07	-	3.127e-06	-
42	1.635e-07	-	-	-	2.428e-07	5.157e-06
43	-	-	1.324e-07	-	-	3.127e-06
44	7.741e-08	-	-	-	-	2.428e-07
64	-	1.454e-07	-	-	-	-
65	-	1.202e-07	-	-	-	-
66	3.220e-09	2.456e-07	1.454e-07	-	-	-
67	4.382e-08	-	3.831e-07	-	-	-
68	-	-	2.456e-07	-	-	-
74	2.749e-07	4.511e-07	-	8.555e-21	-	-
75	-	-	1.142e-06	-	1.223e-20	-
76	7.521e-08	-	4.511e-07	-	8.555e-21	2.153e-21
77	-	-	-	-	-	1.223e-20
78	-	-	-	-	-	8.555e-21
100	2.826e-10	1.447e-08	-	-	-	-

An Analysis of δ-Step Ahead Adaptive Prediction and its Application in Active Noise Control

Xuan Kong and Sen M. Kuo
Department of Electrical Engineering
Northern Illinois University
DeKalb, IL 60115

Abstract This paper discusses the performance of adaptive algorithms in predicting a time series δ-step ahead of available observations. An important application of this is the feed-forward active noise control (ANC) system. When the acoustic/electric delays in the noise canceling sub-system exceeds the acoustic delay of the primary path traveled by the noise, it is necessary to predict the noise δ-step ahead. We present a performance analysis of the feed-forward ANC system under such a condition. The bandwidth of the noise controllable by the ANC system decreases as δ increases. Analysis also shows that noise with sharper spectrum peaks (narrower bandwidth) can be more effectively canceled out by the ANC system with a given δ.

1 Introduction

In a broadband acoustic active noise control (ANC) system [1] as illustrated in Figure 1, the primary noise is sensed by a reference microphone close to the noise source. An adaptive filter uses this reference signal to generate a canceling signal, which is fed to a secondary loudspeaker to cancel the primary noise. An error microphone measures the residual noise and uses it to update the coefficients of the adaptive filter. The filtered-X least-mean-square (LMS) algorithm ([2]–[4]) is a commonly used algorithm for updating the coefficients of the adaptive filter.

The acoustic delay from the reference microphone to the secondary loudspeaker is proportional to the distance from the reference sensor to the secondary source. The electrical delay is the sum of group delay of the digital filter and the total system delay in the anti-aliasing filter, A/D converter, D/A converter, reconstruction filter, loudspeaker, plus the processing time (one sampling period). When the electrical delay is longer than the acoustic delay, the optimal response of the adaptive filter becomes non-causal, and hence unrealizable. Therefore, to attenuate broadband noise, the standard approach to ANC in ducts is to use a long duct. However, installation constraints of commercial systems usually limit the length of duct that can be used. Therefore an in-depth study of the performance of the ANC system with a short duct length is needed.

Feed-forward ANC system functions as adaptive system identification, thus able to control both broadband and narrow-band noises [1]. This paper shows that when the causality constraint is violated, the system is equivalent to an adaptive predictor and can cancel band-limited noise only. In Section 2, we present a mathematical model of the feed-forward ANC system setup suitable for the study of the effects of non-causality. Section 3 details our analysis of the performance of the ANC system under different degree of non-causality and signal conditions. Computer simulations to illustrate the analysis results are presented in Section 4. Some conclusion remarks are given in Section 5.

2 Feed-forward ANC Systems and Causality Constraint

The block diagram of a single-channel feed-forward ANC system is illustrated in Figure 2. $x[n]$ is the reference signal acquired from the reference sensor and $e[n]$ is the signal detected by the error sensor. $P(z)$ is the model for the primary path between the reference sensor and error sensor, while $S(z)$ is the secondary path between the secondary source and the error sensor. An estimate of the secondary path is denoted as $\hat{S}(z)$. The adaptive filter $W(z)$ is updated by the filtered-X LMS algorithm. The z-domain expression for the error signal $e[n]$ is

$$E(z) = X(z)P(z) - X(z)W(z)S(z). \tag{1}$$

To have the residual error to be zero requires $W(z)$ to satisfy

$$W_{opt}(z) = \frac{P(z)}{S(z)}. \tag{2}$$

Suppose $\hat{S}(z) = S(z)$ and the adaptive gain used in the LMS algorithm is sufficiently small so that $W(z)$ changes slowly and the systems $W(z)$ and $S(z)$ can be commuted [1]. To allow a clear understanding of the causality issue, the transfer functions $P(z)$ and $S(z)$ are simplified as pure delays, $z^{-\Delta P}$ and $z^{-\Delta S}$, respectively. When ΔP is larger than ΔS, causality constraint is satisfied. The z-domain error signal given in (1) can be expressed as

$$E(z) = X(z)z^{-\Delta S}[z^{-\delta} - W(z)], \tag{3}$$

where $\delta = \Delta P - \Delta S > 0$. Therefore the ANC system acts like a system identification scheme. When ΔS is larger than ΔP, the causality constraint is violated as we must have

$$E(z) = X(z)z^{-\Delta P}[1 - z^{-\delta}W(z)], \tag{4}$$

where $\delta = \Delta S - \Delta P > 0$. The system is identical to an adaptive predictor scheme shown in Figure 3. Clearly, for the same degree of non-causality (i.e., δ), the adaptive filter $W(z)$ can make a better prediction of the signal $x[n]$ with a narrower bandwidth. For a given bandwidth of $x[n]$, the larger the delay δ is, the more difficult it is for the adaptive algorithm to predict accurately the primary noise $d[n]$. In the next section, we formalize the above intuitive observations via an analysis of the adaptive algorithm.

3 Performance Analysis

In this section, we analyze the performance of the ANC system outlined in the previous section when the causality constraint is violated. Investigation is carried out under a simplified primary path condition to clearly demonstrate the effects of non-causality on the performance of the ANC.

The broadband noise source is commonly modeled by an AR process. A general AR model transfer function can be decomposed into a summation of the second order processes:

$$\sum_{q=1}^{Q} \frac{a_q + b_q z^{-1}}{(1 - \rho_q e^{j\theta_q} z^{-1})(1 - \rho_q e^{-j\theta_q} z^{-1})}, \tag{5}$$

where $\rho_q < 1$ to ensure that the signal $x[n]$ is stationary. In the following analysis, we concentrate on the ANC systems with a reference signal $x[n]$ which is modeled by the following second-order AR process:

$$x[n] = \frac{1}{(1 - \rho e^{j\theta} z^{-1})(1 - \rho e^{-j\theta} z^{-1})} \varepsilon[n]. \tag{6}$$

Since $\varepsilon[n]$ is white, spectrum of the reference signal $x[n]$ peaks at the digital frequency θ. The dependence of the spectrum bandwidth on the parameter ρ is given as follows:

$$\text{Bandwidth} = 2\cos^{-1}\left(\frac{\rho^2 - 4\rho + 1}{2\rho}\right) \tag{7}$$

which decreases to zero as ρ approaches 1. Thus, the closer ρ is to the unit circle, the narrower the bandwidth of $x[n]$ will be. Intuitively, a signal with a narrower bandwidth can be better predicted. One would then expect that when the causality constraint is violated, the performance of a given ANC system will be better for the reference signal $x[n]$ whose model parameter ρ is closer to one. We prove this in the following analysis.

3.1 Moving-Average Interpretation

Before we proceed to the analysis based on the AR-model, let us study the problem using its moving average (MA) equivalent interpretation.

The desired signal for the adaptive filter of the ANC system can be expressed as

$$d[n] = x[n] = \sum_{k=0}^{\infty} w[k]\varepsilon[n-k] \tag{8}$$

with a corresponding adaptive filter output as

$$y[n] = \sum_{k=0}^{\infty} \hat{w}_n[k]\varepsilon[n-k-\delta]. \tag{9}$$

The adaptive algorithm is given by

$$\hat{\mathbf{w}}_{n+1} = \hat{\mathbf{w}}_n + \mu e[n]\underline{\varepsilon}_{n-\delta} \tag{10}$$

where $e[n] = d[n] - y[n]$ and

$$\hat{\mathbf{w}}_n = [\hat{w}_n[0]\ \hat{w}_n[1]\ \hat{w}_n[2]\ \cdots]^T$$
$$\underline{\varepsilon}_n = [\varepsilon[n]\ \varepsilon[n-1]\ \varepsilon[n-2]\ \cdots]^T.$$

Upon convergence, we have the minimum mean square error (MSE) $\xi_{min}(\rho,\delta)$ as

$$\xi_{min}(\rho,\delta) = E[(d[n] - y[n])^2] = \sigma_\varepsilon^2 \sum_{k=0}^{\delta-1} w^2[k]. \tag{11}$$

Obviously, $\xi_{min}(\delta)$ decreases as δ decreases.

We define the noise canceling efficiency η as follows

$$\eta = 1 - \xi_{min}/\xi_{total} \tag{12}$$

to measure the effectiveness of the ANC system. The efficiency η reflects the percentage of the energy in the primary noise which the ANC system is able to reduce, thus it gives us a simple but quantitative means to measure the performance of the ANC system.

The total primary noise energy can be found easily as follows

$$\xi_{total} = E[d^2(n)] = \sigma_\varepsilon^2 \sum_{k=0}^{\infty} w_k^2. \tag{13}$$

Substituting (11) and (13) into (12), we have

$$\eta(\delta) = 1 - \frac{\sigma_\varepsilon^2 \sum_{k=0}^{\delta-1} w_k^2}{\sigma_\varepsilon^2 \sum_{k=0}^{\infty} w_k^2} = \frac{\sum_{k=\delta}^{\infty} w_k^2}{\sum_{k=0}^{\infty} w_k^2}. \tag{14}$$

Thus we conclude that the ANC system efficiency becomes lower as the degree of non-causality (δ) becomes higher.

3.2 Auto-Regressive Model Interpretation

Converting the AR model (6) into the MA model (8), we have

$$x[n] = \frac{1}{(1-\rho e^{j\theta}z^{-1})(1-\rho e^{-j\theta}z^{-1})}\varepsilon[n]$$

$$= \sum_{k=0}^{\infty} \rho^k \frac{\sin(k+1)\theta}{\sin\theta}\varepsilon[n-k]. \tag{15}$$

Substituting

$$w_k = \rho^k \sin[(k+1)\theta]/\sin\theta$$

into (14) we have

$$\eta(\rho,\delta) = \frac{\sum_{k=\delta}^{\infty} \rho^{2k}[1-\cos(2k+1)\theta]}{\sum_{k=0}^{\infty} \rho^{2k}[1-\cos(2k+1)\theta]}. \tag{16}$$

Fixing ρ, we can easily show that ANC efficiency $\eta(\rho,\delta)$ increases as the degree of non-causality decreases. We can also show that for any fixed δ_0, $\eta(\rho,\delta_0)$ increases monotonely with ρ. This fact, together with $\eta(0,\delta_0) = 0$ and $\eta(1,\delta_0) = 1$, shows that the narrower the bandwidth of the reference noise is, the higher the efficiency of the ANC system becomes.

A direct analysis which only relies on the AR representation of the signal $x[n]$ is also possible. The adaptive algorithm directly updates the model parameters as follows:

$$\hat{\mathbf{w}}_{n+1} = \hat{\mathbf{w}}_n + \mu(x[n] - y[n])\mathbf{x}_{n-\delta} \tag{17}$$

with

$$y[n] = \hat{\mathbf{w}}_n^T \mathbf{x}_{n-\delta} = \sum_{m=0}^{M-1} \hat{w}_m[n]x[n-m-\delta]$$

and

$$\mathbf{x}_{n-\delta} = [x[n-\delta]\ x[n-\delta-1]\ \cdots\ x[n-\delta-M+1]]^T.$$

The optimal weight vector $\hat{\mathbf{w}}_{opt}$ is given by

$$\hat{\mathbf{w}}_{opt} = \mathbf{R}_x^{-1}\hat{\mathbf{r}}_\delta \tag{18}$$

where $\mathbf{R}_x$ is a Toeplitz matrix generated by $(\gamma_x[0]\ \gamma_x[1]\ \cdots \gamma_x[M-1])^T$ and $\hat{\mathbf{r}}_\delta = (\gamma_x[\delta]\ \gamma_x[\delta+1]\ \cdots \gamma_x[\delta+M-1])^T$ with $\gamma_x[l] = E(x[n]x[n+l])$.

Define $\tilde{\mathbf{w}}_n = \hat{\mathbf{w}}_n - \hat{\mathbf{w}}_{opt}$, we have from (17)

$$\tilde{\mathbf{w}}_{n+1} = \tilde{\mathbf{w}}_n + \mu(x[n] - \tilde{\mathbf{w}}_n^T \mathbf{x}_{n-\delta} - \hat{\mathbf{w}}_{opt}^T \mathbf{x}_{n-\delta})\mathbf{x}_{n-\delta}. \tag{19}$$

It was established in [6] that the trajectory of (19) can be made arbitrarily close to the trajectory of the following averaged system by reducing μ

$$\bar{\mathbf{w}}_{n+1} = (\mathbf{I} - \mu\mathbf{R}_x)\bar{\mathbf{w}}_n. \tag{20}$$

In practice, the adaptive gain μ used is small enough to allow a good approximation of $\tilde{\mathbf{w}}_n$ by $\bar{\mathbf{w}}_n$.

One important conclusion we can draw from (20) is that the speed of the convergence for the weight vector $\tilde{\mathbf{w}}_n$ remains the same regardless of the degree of the non-causality (δ). This is very important as operators of ANC systems would like to know how quickly the adaptation will occur even the accurate knowledge of the degree of non-causality is not available.

If the reference noise $x[n]$ is a first order AR process of the form

$$x[n] = \rho x[n-1] + \varepsilon[n]$$

it leads to

$$\eta(\rho, \delta) = \frac{\rho^{2\delta}}{1-\rho^2} \tilde{\mathbf{r}}_0^T \tilde{\mathbf{R}}_x^{-1} \tilde{\mathbf{r}}_0, \quad \tilde{\mathbf{r}}_\delta = \mathbf{r}_\delta / \sigma_x^2. \tag{21}$$

where $\tilde{\mathbf{R}}_x$ is a Toeplitz matrix generated by $[1\ \rho\ \rho^2\ \cdots\ \rho^{M-1}]^T$. The above equation indicates that when the degree of non-causality increases, the noise canceling efficiency decreases. Also, as ρ increases (but < 1), the efficiency increases.

4 Simulation Results

We present simulation results to verify the theoretical development in Section 3. We use the following reference noise model for all simulations:

$$x[n] = \frac{1}{(1-\rho e^{j\theta} z^{-1})(1-\rho e^{-j\theta} z^{-1})} \varepsilon[n]. \tag{22}$$

where $\varepsilon[n]$ is a white noise with unit variance, $\theta = \pi/6$. The secondary path model is also fixed at

$$S(z) = z^{-\delta}$$

where the delay parameter δ is to be chosen. The simplified primary path model is $P(z) = z^{-\Delta P}$ in general. Since we are primarily concerned in this paper about the non-causality, we will fix $\Delta P = 0$ and vary δ in $S(z) = z^{-\delta}$ to realize different simulation conditions for non-causality.

Figure 4 shows the efficiency of the ANC system ($L = 6$, $\mu = 0.01/\sigma_x^2$) under various conditions. Each curve represents a simulation condition for a specific bandwidth of the reference signal $x[n]$ (via the specification of ρ). Various degrees of non-causality (δ) are tested. For a specific ρ and δ, the estimated variance of the residual error was calculated based on the last 1000 iterations of a total of 5000 iterations. The estimated variances were then averaged over 10 realizations with identical initial conditions and independent realizations of $\varepsilon[n]$. The results confirm the two important conclusions drawn in Section 3: The ANC system efficiency decreases exponentially as a function of the degree of non-causality δ and decreases as the bandwidth of the reference signal increases.

Figure 5 shows that the convergence speed of the ANC system ($L = 6$) is independent of the degree of the non-causality (δ). Each curve represents the averaged results over ten independent realizations of the adaptive algorithm. Indeed the speed of the convergence is the same for different levels of non-causality as measured by δ. The adaptive gain for all instances is set at $\mu = 0.005/\sigma_x^2$.

5 Conclusions

In this paper, we presented an analysis of the feed-forward ANC system regarding its noise canceling performance when the causality constraint is violated. Under such a non-causal signal condition, the feed-forward ANC system acts like a δ-step ahead predictor based on the available reference inputs. We demonstrated that the larger the degree of the non-causality is, the worse the ANC system performance is. The noise canceling results of the ANC is less effective when the noise spectrum peak bandwidth becomes wider.

References

[1] S. M. Kuo and D. R. Morgan, *Active Noise Control Systems – Algorithms and DSP Implementations*, New York: Wiley, 1996.

[2] D. R. Morgan, "An analysis of multiple correlation cancellation loops with a filter in the auxiliary path," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 454–467, 1980.

[3] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1985.

[4] J. C. Burgess, "Active adaptive sound control in a duct: A computer simulation," *J. Acoust. Soc. Am.*, vol. 70, no. 3, pp. 715–725, Sept. 1981.

[5] S. Haykin, *Adaptive Filter Theory*, 3rd Edition, Prentice-Hall, Englewood Cliff, NJ, 1996.

[6] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*, Prentice-Hall, Englewood Cliff, NJ, 1995.

[7] J.P. Burg, "Maximum entropy spectral analysis," in *Proc. 37th Meeting Soc. Exploratory Geophys.*, 1967.

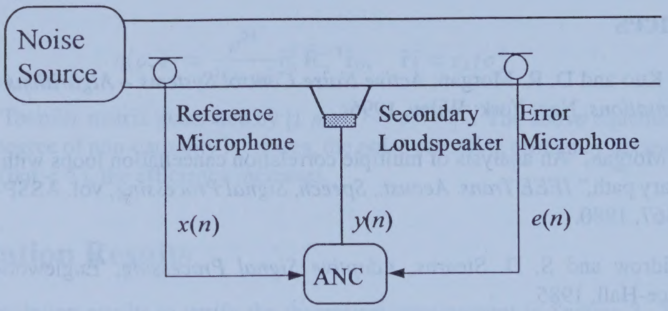

Noise Source
Reference Microphone
Secondary Loudspeaker
Error Microphone
x(n)
y(n)
e(n)
ANC

Figure 1: Acoustic active noise control in ducts.

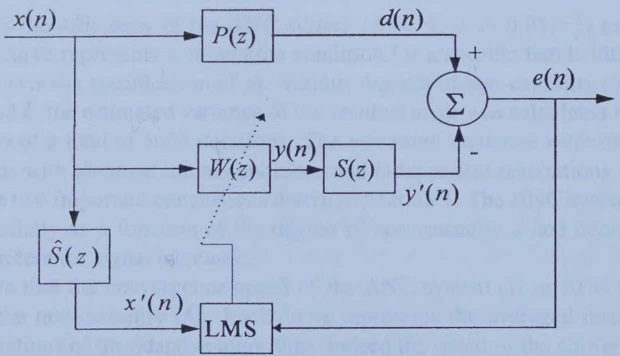

x(n)
P(z)
d(n)
+
e(n)
Σ
y(n)
W(z)
S(z)
-
y'(n)
Ŝ(z)
x'(n)
LMS

Figure 2: Block diagram of ANC system using the filtered-X LMS algorithm

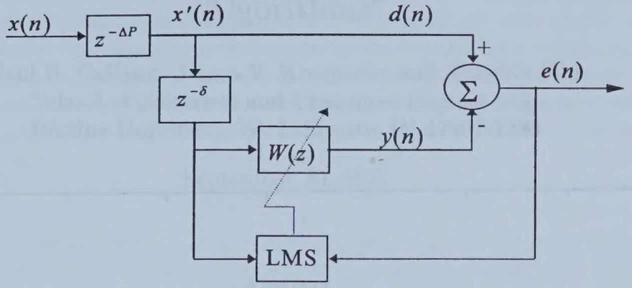


Figure 3: Simplified feed-forward ANC system, acting as an adaptive predictor.

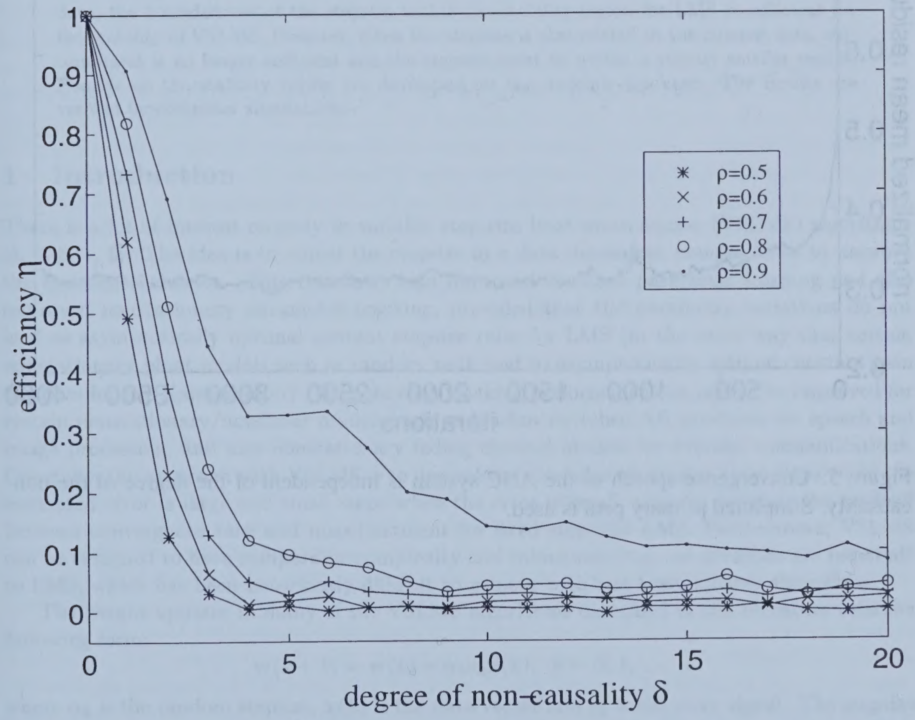


Figure 4: Performance of the ANC system as a function of the bandwidth of the primary noise and the degree of non-causality. Simplified primary path is used.

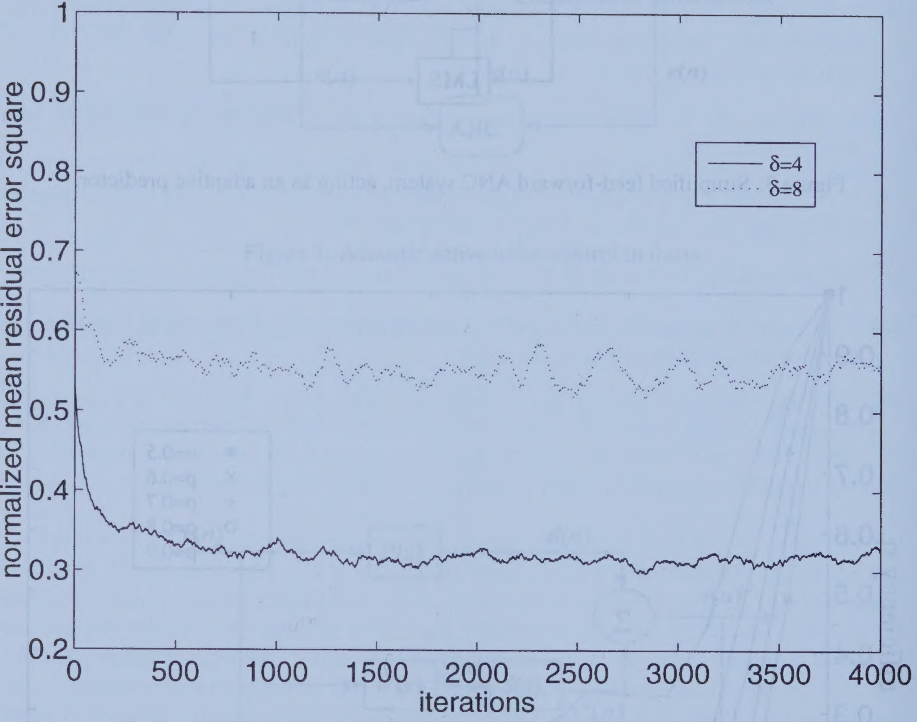


Figure 5: Convergence speech of the ANC system is independent of the degree of the non-causality. Simplified primary path is used.

New Results on Stability of Variable Stepsize LMS Algorithms*

Saul B. Gelfand, James V. Krogmeier and Yongbin Wei
School of Electrical and Computer Engineering,
Purdue University, W. Lafayette, IN 47907-1285

September 30, 1997

Abstract

The stability of variable stepsize LMS (VSLMS) algorithms with uncorrelated stationary Gaussian data is studied. It has sometimes been assumed in the literature that VSLMS algorithms are stable (in the sense of mean-square bounded weights) provided that the stepsize is constrained to lie within the (mean-square) stability region for the conventional fixed-stepsize LMS algorithm. For a single tap filter we find an exact expression for the stability region of VSLMS. It is shown that when the stepsize is determined by the past data, the boundedness of the stepsize within the stability region for LMS is sufficient for the stability of VSLMS. However, when the stepsize is also related to the current data, this constraint is no longer sufficient and the stepsize must lie within a strictly smaller region. Bounds on the stability region are developed for the multiple tap case. The results are verified by computer simulations.

1 Introduction

There is a lot of interest recently in variable step-size least mean square (VSLMS) algorithms [4, 6, 8, 1, 5]. The idea is to adjust the stepsize in a data dependent manner so as to improve the transient behavior. This translates into improved constant parameter learning and also improved nonstationary parameter tracking, provided that the parameter variations do not lead to asymptotically optimal contant stepsize rules for LMS (in the same way that certain nonstationary plant models such as random walk lead to asymptotically optimal constant gain matrices for the Kalman filter). We note that tracking performance can in fact be improved for certain nonstationary/nonlinear models such as Markov switched AR processes for speech and image processing, and also nonstationary fading channel models for wireless communications. Generally, the approach with VSLMS is to devise step size rules which give large steps when the estimated error is large and small steps when the error is small, thereby avoiding the tradeoff between convergence rate and misadjustment for fixed step size LMS. Furthermore, VSLMS can be designed to have comparable complexity and robustness (e.g., no divisions are required) to LMS, which has been notoriously difficult to achieve with fast least squares algorithms.

The weight updates in many of the VSLMS algorithms described in the literature take the following form:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \alpha_k e_k \mathbf{x}(k), \quad k = 0, 1, \ldots$$

where α_k is the random stepsize, $\mathbf{x}(k)$ is the data vector and e_k is the error signal. The stepsize α_k in the literature is typically computed in one of two ways: (i) α_k is a *prior* stepsize which depends on $\mathbf{x}(i), e_i$, for $i < k$, i.e., on the past data and errors; and (ii) α_k is a *posterior* stepsize

---

*This work was partially supported by National Science Foundation Grant 9406073-NCR and a David Ross Fellowship.

which depends on $\mathbf{x}(i), e_i$, for $i \leq k$, i.e., on the current and past data and errors. Furthermore, a bound is enforced on α_k to ensure stability (presumably): $\underline{\alpha} \leq \alpha_k \leq \overline{\alpha}$.

The (classical-type) analysis of these VSLMS algorithms in the literature proceeds in two steps [6, 8](see also [5] for a different approach). First, a rigorous stability analysis is made (or claimed) leading to conditions for mean-square (MS) bounded weights and bounded mean-square error (MSE), and second, an approximate performance analysis is carried out, including convergence to and characterization of the asymptotic weight mean, covariance and MSE. That one can at least guarantee stability (MS bounded weights) rigorously would seem to support the performance analysis. Now for stationary uncorrelated Gaussian data and constant stepsize α it is known that a sufficient condition for MS bounded weights is that [2]

$$0 < \alpha < \frac{2}{3\mathrm{Tr}\{\mathbf{R}\}}$$

where $\mathbf{R} = \mathrm{E}\{\mathbf{x}(k)\mathbf{x}(k)^T\}$ (we deal with real data case here for simplicity). For the same data model but now with variable stepsize α_k, it is tacitly assumed that a sufficient condition for MS bounded weights is that

$$0 < \underline{\alpha} \leq \alpha_k \leq \overline{\alpha} < \frac{2}{3\mathrm{Tr}\{\mathbf{R}\}}. \tag{1}$$

This assumption is straightforwardly correct if α_k is deterministic, but the usual case is where the α_k is data-dependent. One wonders, for example, if the gap between $\underline{\alpha}$ and $\overline{\alpha}$ plays a role in stability for such random stepsizes.

In this paper we give both necessary and sufficient conditions on the stepsize sequence for the stability (MS bounded weights) of VSLMS in the classical setting of stationary uncorrelated Gaussian data. For the special case of a single tap filter, an exact characterization of the stability region is obtained, whereby it is shown that (1) is a sufficient condition for MS bounded weights for prior stepsize rules (which only depend on past data), but is *not* a sufficient condition for posterior step size rules (which also depend on the current data). We also present an explicit example of unstable behavior in the latter case when (1) is satisfied. For multiple tap filters some bounds on the stability region are obtained. This work is a generalization of the results in [2] for stability of fixed step size LMS algorithms, and is important for rigorously establishing (at least under strong classical assumptions on the data) the stability of a class of adaptive algorithms which are of increasing interest in practical applications. It complements the approximate analysis of the VSLMS in [6, 8] which was shown to have value for the prediction of steady state behavior.

1.1 Some comments on the method and goals of the analysis

The asymptotic analysis of the least mean squares (LMS) fixed stepsize stochastic gradient algorithm for adaptive filtering can be viewed as having classical and modern phases. In the classical approach, LMS is analyzed under strong conditions such stationary uncorrelated Gaussian data [2] or stationary $M-$ dependent data [7]. In the modern approach LMS is analyzed under weaker conditions on the data, e.g., as an output of a uniformly asymptotically stable linear system driven by white noise [3]. The key results in these works is demonstrating the p -th mean exponential stability of the unforced equation. If this exponential stability holds for large enough p, then one has MS bounded weights (for stationary uncorrelated Gaussian data it is enough to consider $p = 2$, i.e., MS exponential stability). Also, a convergence and performance analysis including asymptotic weight mean, covariance and also MSE for learning and tracking certain models of parameter variation can be obtained (and justified). There are also many works which circumvent the significant difficulties associated with demonstrating exponential stability for unbounded inputs, including assuming bounded data and/or bounded weights (see the recent book by [9] for a current list of references). In the case of bounded data, the exponential stability is then obtained more easily, and in the case of bounded weights the convergence and tracking behavior can be directly obtained by characterizing limit points and distributions. Such approaches may be the only viable approach for rigorous performance analysis of LMS (and more complex adaptive algorithms) with general dependent data models.

The classical approach to analyzing LMS is constructive, and is primarily concerned with conditions on the stepsize which ensure exponential stability (for stationary uncorrelated Gaussian data a characterization of the stepsize has actually been achieved[2]). The modern approach is more qualitative and is primarily concerned with conditions on the data which ensures exponential stability for some small enough stepsize (a characterization in terms of persistance of excitiation condition has recently been obtained [3]). Furthermore, we observe that approaches which assume bounded data and/or bounded weights seem most useful for performance analysis. Indeed, although the assumption of bounded data may still require additional conditions on the stepsize for exponential stability, such conditions typically depend on the specific bound on the data, which for realistic data bounds results in overly conservative stepsizes.

The upshot of this discussion is that in terms of predicting conditions on stepsize which guarantee stability, the classical approach with (say) the assumption of stationary uncorrelated Gaussian data gives concrete results, and is well-known to have empirical validity for more general data models. In this paper we significantly extend this approach to treat the stability of the variable step-size LMS (VSLMS) algorithm.

2 The Set-up

Consider the VSLMS algorithm in the form

$$\begin{aligned} \mathbf{w}(k+1) &= (\mathbf{I} - \alpha_k \mathbf{x}(k)\mathbf{x}(k)^T)\mathbf{w}(k) + \alpha_k n_k \mathbf{x}(k) \\ e_k &= (\mathbf{w}^* - \mathbf{w}(k))^T \mathbf{x}(k) + n_k \end{aligned} \tag{2}$$

for $k = 0, 1, \ldots$ We shall assume that the regressors $\{\mathbf{x}(k)\}$are $d-$dimensional stationary independent Gaussian vectors with mean $\mathbf{0}$ and covariance $\mathbf{R}$ (postitive definite, with positive eigenvalues $\lambda_1, \ldots, \lambda_d$), the noise $\{n_k\}$ is stationary independent sequence with mean 0 and variance σ_n^2 , and $\{\mathbf{x}(k)\}$, $\{n_k\}$ are independent [1]. Let Y_k be the $\sigma-$ field generated by $\mathbf{x}(i), e_i$, for $i \leq k$. We shall consider prior and posterior step-sizes α_k which are Y_{k-1} and Y_k measurable, respectively.

We want to examine how the stability of the VSLMS algorithm depends on the step size sequence under the above assumptions on the data. As mentioned above, for LMS (with fixed step size), the key thing to show is the MS exponential stabilty of the weight vector equation, which implies MS bounded weights. However, for VSLMS with data dependent step-size, it is not obvious that the MS exponential stability implies MS bounded weights. For this reason, we directly examine the MS boundedness of the weights , being careful not to truncate terms which depend implicitly on the weights through the step size. The approach we shall take is to upper bound the diagonal elements of the weight error correlation matrix (i.e. essentially the weight variances) by the solution of a (time-invariant) asymptotically stable linear difference equation, generalizing the method in [2] for LMS.

Define the *Mean Square (MS) stability region*

$$\begin{aligned} S &= \{(\underline{\alpha}, \overline{\alpha}) : \sup_k \mathrm{E}\{|\mathbf{w}(k)|^2\} < \infty \text{ for all } \{\alpha_k\} \\ &\qquad \text{such that } 0 < \underline{\alpha} \leq \alpha_k \leq \overline{\alpha} \text{ w.p. } 1\}. \end{aligned}$$

From [2] for the fixed stepsize case $\alpha_k = \alpha$ it is known that the weights are MS bounded (with bound independent of the initial weight values) if and only if $0 < \alpha < \alpha^*$, where

---

[1] this system identification model is slightly more general than the adaptive filtering model of stationary independent Gaussian data (data vectors and desired signals)

$$\alpha^* = \sup\left\{ \alpha : 0 < \alpha < \frac{1}{\lambda_i},\ i = 1, ..., d;\ \sum_{i=1}^{d} \frac{\alpha\lambda_i}{2(1-\alpha\lambda_i)} < 1 \right\}.$$

It follows that $S \subset \{(\underline{\alpha}, \overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*\}$. Now define the *lower envelope* of the MS stability region

$$\underline{\alpha}(\overline{\alpha}) = \inf\{\underline{\alpha} : (\underline{\alpha}, \overline{\alpha}) \in S\},\quad 0 < \overline{\alpha} < \alpha^*.$$

($\underline{\alpha}(\overline{\alpha})$ is well-defined since $(\overline{\alpha}, \overline{\alpha}) \in S$). Observe that

- if $0 < \overline{\alpha} < \alpha^*$ and $\underline{\alpha}(\overline{\alpha}) < \underline{\alpha}$, then $\mathrm{E}\{|\mathbf{w}(k)|^2\}$ is bounded for all $\{\alpha_k\}$ such that $\underline{\alpha} \leq \alpha_k \leq \overline{\alpha}$ w.p. 1.
- if $0 < \overline{\alpha} < \alpha^*$ and $\underline{\alpha} < \underline{\alpha}(\overline{\alpha})$, then $\mathrm{E}\{|\mathbf{w}(k)|^2\}$ is unbounded for some $\{\alpha_k\}$ such that $\underline{\alpha} \leq \alpha_k \leq \overline{\alpha}$ w.p. 1.

Our goal is to characterize as tightly as possible the MS stability region and lower envelope. We first treat the single tap ($d = 1$) case, since it is rather special (and far simpler), and we can give the exact stability region in this case.

3 Stability Region for Single Tap Filter

For the case where $d = 1$, we have the following characterizations of the MS stability region for the VSLMS algorithm in (2). Here $\lambda = \mathrm{E}\{x(k)^2\}$ and $\alpha^* = \frac{2}{3\lambda}$.

Proposition 1 *If α_k depends on Y_{k-1} then*

$$S = \{(\underline{\alpha}, \overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*\}.$$

Proposition 2 *If α_k depends on Y_k then*

$$S = \{(\underline{\alpha}, \overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*,\ \rho(\underline{\alpha}, \overline{\alpha}) < 1\}.$$

where

$$\begin{aligned} \rho(\underline{\alpha}, \overline{\alpha}) &= 1 - 2\overline{\alpha}\lambda + 3\overline{\alpha}^2\lambda^2 + 4(\overline{\alpha} - \underline{\alpha})\lambda\Phi_1\left(\frac{1}{\sqrt{2p\lambda}}\right) \\ &\quad - 4(\overline{\alpha}^2 - \underline{\alpha}^2)\lambda^2\Phi_2\left(\frac{1}{\sqrt{2p\lambda}}\right) \qquad (3) \\ \Phi_n(x) &= \frac{2}{\sqrt{\pi}}\int_0^x t^{2n}e^{-t^2}\,dt \end{aligned}$$

and $p = (\underline{\alpha} + \overline{\alpha})/2$. Furthermore, if $\rho(\underline{\alpha}, \overline{\alpha}) \geq 1$ then an example which has unbounded weight variance is

$$\alpha_k = \left\{ \begin{array}{ll} \overline{\alpha} & \text{if } px^2(k) > 1 \\ \underline{\alpha} & \text{if } px^2(k) \leq 1 \end{array} \right. . \qquad (4)$$

Remark 1 *It can be shown using the implicit function theorem that the lower envelope for the posterior step size is the unique solution of $\rho(\underline{\alpha}(\overline{\alpha}), \overline{\alpha}) = 1$, $0 < \overline{\alpha} < \alpha^*$, and is smooth (at least C^2) and strictly convex. Also, it is easy to show that $\underline{\alpha}(\overline{\alpha}) > 0$ for $0 < \overline{\alpha} < \alpha^*$. Hence there is a* nontrivial *(i.e., not identically zero) lower envelope for the posterior stepsize, and the stability region is strictly smaller than for deterministic (or prior) stepsize.*

Next we give the proofs of these results.

Proof of Proposition 1. Let $v(k) = w(k) - w^*$, and

$$f(\alpha) = \mathrm{E}\left\{\left(1 - \alpha x(k)^2\right)^2\right\} = 1 - 2\alpha\lambda + 3\alpha^2\lambda^2.$$

Then we can write

$$\begin{aligned} \mathrm{E}\left\{v(k+1)^2\right\} &= \mathrm{E}\left\{f(\alpha_k)v(k)^2\right\} + \mathrm{E}\left\{\alpha_k^2\right\}\lambda\sigma_n^2 \\ &\leq \rho(\underline{\alpha},\overline{\alpha})\mathrm{E}\left\{v(k)^2\right\} + c(\overline{\alpha}), \end{aligned} \tag{5}$$

where

$$\rho(\underline{\alpha},\overline{\alpha}) = \max_{\alpha\in[\underline{\alpha},\overline{\alpha}]} f(\alpha)$$

and $c(\overline{\alpha}) = \overline{\alpha}^2\lambda\sigma_n^2 > 0$. Now it is seen that $f(\alpha)$ is maximized at

$$\alpha_{opt} = \begin{cases} \overline{\alpha} & \text{if } 3p\lambda > 1, \\ \underline{\alpha} & \text{if } 3p\lambda \leq 1, \end{cases}$$

and so

$$\rho(\underline{\alpha},\overline{\alpha}) = f(\alpha_{opt}) = \begin{cases} 1 - 2\overline{\alpha}\lambda + 3\overline{\alpha}^2\lambda^2 & \text{if } 3p\lambda > 1, \\ 1 - 2\underline{\alpha}\lambda + 3\underline{\alpha}^2\lambda^2 & \text{if } 3p\lambda \leq 1. \end{cases}$$

Since $\rho(\underline{\alpha},\overline{\alpha}) > 0$ we can replace the upper bound in (5) with equality to get

$$\xi(k+1) = \rho(\underline{\alpha},\overline{\alpha})\xi(k) + c(\overline{\alpha}); \quad \xi(0) = \mathrm{E}\left\{v(0)^2\right\} \tag{6}$$

and $\mathrm{E}\left\{v(k)^2\right\} \leq \xi(k)$ for all k. Now (6) is exponentially stable, and consequently $\sup_k \mathrm{E}\left\{v(k)^2\right\} \leq \sup_k \xi(k) < \infty$, if and only if $\rho(\underline{\alpha},\overline{\alpha}) < 1$. Hence $S \supset \{(\underline{\alpha},\overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*, \rho(\underline{\alpha},\overline{\alpha}) < 1\}$ and we already know that $S \subset \{(\underline{\alpha},\overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*\}$. It is easily checked that if $0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*$ then $\rho(\underline{\alpha},\overline{\alpha}) < 1$. The proposition follows. □

Proof of Proposition 2. Let $v(k) = w(k) - w^*$, and

$$f(\alpha, x) = \left(1 - \alpha x^2\right)^2 = 1 - 2\alpha x^2 + \alpha^2 x^4.$$

Then using Holder's inequality we have

$$\begin{aligned} \mathrm{E}\left\{v(k+1)^2\right\} &= \mathrm{E}\left\{f(\alpha_k)v(k)^2\right\} + \mathrm{E}\left\{2f\left(\alpha_k, x(k)\right)^{\frac{1}{2}} v(k)\alpha_k x(k) n_k\right\} \\ &\quad + \mathrm{E}\left\{\alpha_k^2\right\}\lambda\sigma_n^2 \\ &\leq \rho(\underline{\alpha},\overline{\alpha})\mathrm{E}\left\{v(k)^2\right\} + 2\left(\rho(\underline{\alpha},\overline{\alpha})c(\overline{\alpha})\right)^{\frac{1}{2}}\left(\mathrm{E}\left\{v(k)^2\right\}\right)^{\frac{1}{2}} + c(\overline{\alpha}) \end{aligned} \tag{7}$$

where

$$\rho(\underline{\alpha},\overline{\alpha}) = \mathrm{E}\left\{\max_{\alpha\in[\underline{\alpha},\overline{\alpha}]} f(\alpha, x(k))\right\},$$

and $c(\overline{\alpha}) = \overline{\alpha}^2\lambda\sigma_n^2 > 0$. Now it is seen that $f(\alpha, x(k))$ is maximized at

$$\alpha_{k,opt} = \begin{cases} \overline{\alpha} & \text{if } px(k)^2 > 1, \\ \underline{\alpha} & \text{if } px(k)^2 \leq 1, \end{cases}$$

and so

$$\begin{aligned} \rho(\underline{\alpha},\overline{\alpha}) &= \mathrm{E}\left\{f(\alpha_{k,opt} x(k))\right\} \\ &= \mathrm{E}\left\{\left(1 - 2\overline{\alpha}x(k)^2 + \overline{\alpha}^2 x(k)^4\right) I\left(px(k)^2 > 1\right)\right\} \\ &\quad + \mathrm{E}\left\{\left(1 - 2\underline{\alpha}x(k)^2 + \underline{\alpha}^2 x(k)^4\right) I\left(px(k)^2 < 1\right)\right\} \end{aligned}$$

where $I(\cdot)$ is an indicator function. This expression is readily evaluated to the one given in (3).

Since $\rho(\underline{\alpha}, \overline{\alpha}) > 0$ we can replace the upper bound in (7) with equality to get

$$\xi(k+1) = \rho(\underline{\alpha}, \overline{\alpha})\xi(k) + O\left(\xi(k)^{\frac{1}{2}}\right) + c(\overline{\alpha}); \quad \xi(0) = \mathrm{E}\left\{v(0)^2\right\} \tag{8}$$

and $\mathrm{E}\left\{v(k)^2\right\} \leq \xi(k)$ for all k. Now (8) has an exponentially stable compact invariant set, and consequently $\sup_k \mathrm{E}\left\{v(k)^2\right\} \leq \sup_k \xi(k) < \infty$, if and only if $\rho(\underline{\alpha}, \overline{\alpha}) < 1$. Hence $S \supset \{(\underline{\alpha}, \overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*, \rho(\underline{\alpha}, \overline{\alpha}) < 1\}$. and we already know that $S \subset \{(\underline{\alpha}, \overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*\}$. Now suppose $\rho(\underline{\alpha}, \overline{\alpha}) \geq 1$ and let $\alpha_k = \alpha_{k,opt}$. Then it is easy to see that

$$\mathrm{E}\left\{v(k+1)^2\right\} \geq \rho(\underline{\alpha}, \overline{\alpha})\mathrm{E}\left\{v(k)^2\right\} + c(\underline{\alpha}),$$

and so $\mathrm{E}\left\{v(k)^2\right\} \to \infty$. The proposition follows. □

3.1 An Example

In Fig. 1 we plot the the lower envelope of the MS stability region (the stability region lies above the envelope and below the line $\underline{\alpha} = \overline{\alpha}$) for posterior stepsize which depend on current and past data and errors (Proposition 2 above). Here $\lambda = 1$ and consequently $\alpha^* = \frac{2}{3}$. Observe that the MS stability region for posterior stepsize is strictly smaller than for deterministic stepsize (which coincides with prior stepsize). This indeed shows that (1) is not a sufficient condition for MS bounded weights even with stepsize which depends only on past and current data. We compare this result with simulation of VSLMS for $w^* = 1$, $\sigma_n^2 = 0.01$, and the stepsize given in Eq. (4). We set $\overline{\alpha}$ to $\frac{1}{3}$ so that from Fig. 1 the smallest value of $\underline{\alpha}$ for MS bounded weights (and hence bounded MSE) is about 0.016. In Figs. 2 and 3 we show the MSE learning curve (averaged over 100 trials) for VSLMS with $\underline{\alpha} = 0.0032$ and $\underline{\alpha} = 0.08$, respectively. We also show the MSE learning curve of LMS with $\alpha = \frac{1}{3}$. The MSE of VSLMS in Fig. 2 appears to diverge, while that in Fig. 3 appears to be bounded, in accord with the theory.

4 Stability Region for Multiple Tap Filter

For the case where $d > 1$, we have the following approximations of the MS stability region for the VSLMS algorithm in (2).

Proposition 3 *If α_k depends on Y_{k-1} then*

$$\begin{aligned} S \;\supset\; & \{(\underline{\alpha}, \overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*; \\ & \sum_{i=1}^{d} \frac{\overline{\alpha}^2 \lambda_i}{\min\left[2\overline{\alpha}\left(1 - \overline{\alpha}\lambda_i\right), 2\underline{\alpha}\left(1 - \underline{\alpha}\lambda_i\right) + \left(\overline{\alpha}^2 - \underline{\alpha}^2\right)\lambda_i\right]} < 1\}. \end{aligned} \tag{9}$$

Proposition 4 *If α_k depends on Y_k then*

$$\begin{aligned} S \;\supset\; & \{(\underline{\alpha}, \overline{\alpha}) : 0 < \underline{\alpha} \leq \overline{\alpha} < \alpha^*; \ \rho_i(\underline{\alpha}, \overline{\alpha}) < 1, \ i = 1, ..., d; \\ & \sum_{i=1}^{d} \frac{\left(\overline{\alpha}^2 + \beta_1\right)\lambda_i^2}{1 - \rho_i(\underline{\alpha}, \overline{\alpha}) + \left(\overline{\alpha}^2 + \beta_1\right)\lambda_i^2 + \beta_2\lambda_i} < 1\}, \end{aligned} \tag{10}$$

where

$$\begin{aligned}\rho_i(\underline{\alpha},\overline{\alpha}) &= 1-2\overline{\alpha}\lambda_i+3\overline{\alpha}^2\lambda_i^2+(\overline{\alpha}-\underline{\alpha})\lambda_i\left(4\Phi_1\left(\frac{1}{\sqrt{2p\lambda_i}}\right)+\frac{1}{4\pi}(d-1)\right)\\ &\quad -(\overline{\alpha}^2-\underline{\alpha}^2)\lambda_i^2\left(4\Phi_2\left(\frac{1}{\sqrt{2p\lambda_i}}\right)-\frac{1}{\pi}(d-1)\right),\end{aligned}$$

$$\begin{aligned}\beta_1 &= \frac{1}{4\pi}\left(\overline{\alpha}^2-\underline{\alpha}^2\right)(d-1),\\ \beta_2 &= \frac{1}{4\pi}\left(\overline{\alpha}-\underline{\alpha}\right),\end{aligned}$$

$$\Phi_n(x)=\frac{2}{\sqrt{\pi}}\int_0^x t^{2n}e^{-t^2}dt,$$

and $p=(\underline{\alpha}+\overline{\alpha})/2$.

Remark 2 *The set inclusions in (9) and (10) implicitly specify upper bounds* $\underline{\alpha}^u(\overline{\alpha})\geq\underline{\alpha}(\overline{\alpha})$ *on the lower envelope for the prior and posterior step-size cases, respectively. It is not known whether a nontrivial lower envelope exists for prior or posterior step size rules when* $d>1$

The proofs of Propositions 3 and 4 are rather involved and will appear elsewhere. In Figures 4 and 5. we show some plots of the guaranteed stability region for $d=2$ and various eigenvalue spreads

References

[1] J. Evans, P. Xue, and B. Liu. Analysis and implementation of variable step size adaptive algorithms. *IEEE Transactions on Signal Processing*, 41(8):2517–2534, August 1993.

[2] A. Feuer and E. Weinstein. Convergence analysis of LMS filters with uncorrelated Gaussian data. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(1):222–230, February 1985.

[3] L. Guo, L. Ljung, and G. Wang. Necessary and sufficient conditions for stability of LMS. *IEEE Transactions on Automatic Control*, 42(6):761–770, June 1997.

[4] R. Harris, D. Chabries, and F. A. Bishop. A variable step (VS) adaptive filter algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(2):309–316, April 1986.

[5] H. Kushner and J. Yang. Analysis of adaptive step-size sa algorithms for parameter tracking. *IEEE Transactions on Automatic Control*, 40(8):1403–1410, August 1995.

[6] R. Kwong and E. W. Johnston. A variable step size LMS algorithm. *IEEE Transactions on Signal Processing*, 40(7):1633–1642, July 1992.

[7] O. Macchi. *Adaptive processing : the least mean squares approach with applications in transmission.* John Wiley and Sons, New York, 1995.

[8] V. J. Mathews and Z. Xie. A stochastic gradient adaptive filter with gradient adaptive step size. *IEEE Transactions on Signal Processing*, 41(6):2075–2087, June 1993.

[9] V. Solo and X. Kong. *Adaptive Signal Processing Algorithms: Stability and Performance.* Prentice Hall, Englewood Cliffs, NJ, 1995.

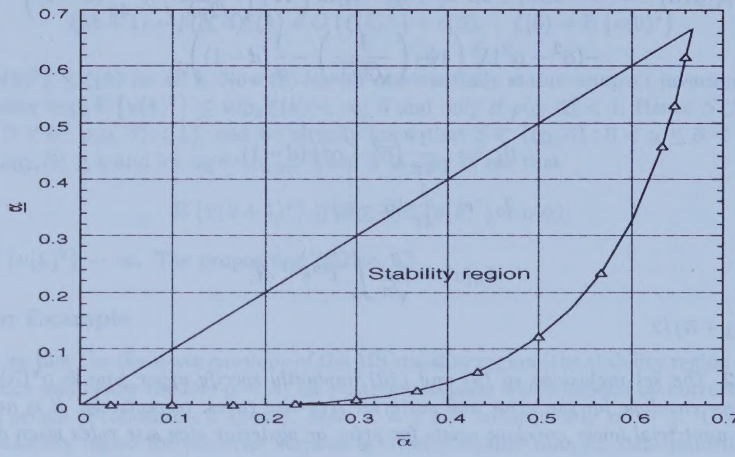


Figure 1: Posterior stepsize stability region for VSLMS, $d = 1$

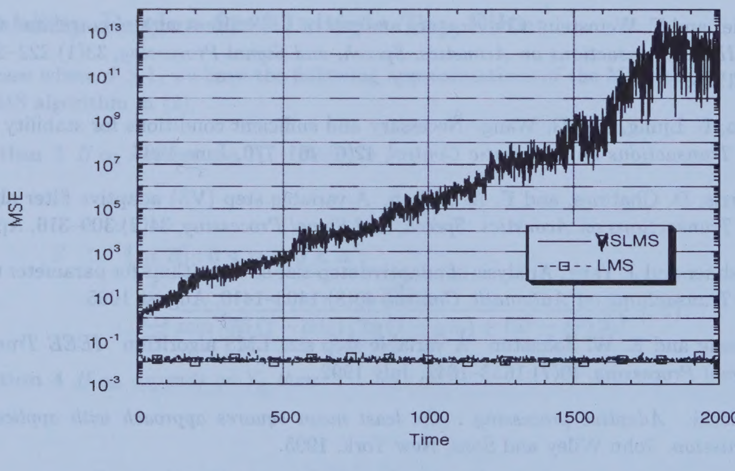


Figure 2: MSE learning curves for LMS with $\alpha = 0.33$ and VSLMS with $\underline{\alpha} = 0.0032$ and $\overline{\alpha} = 0.33$

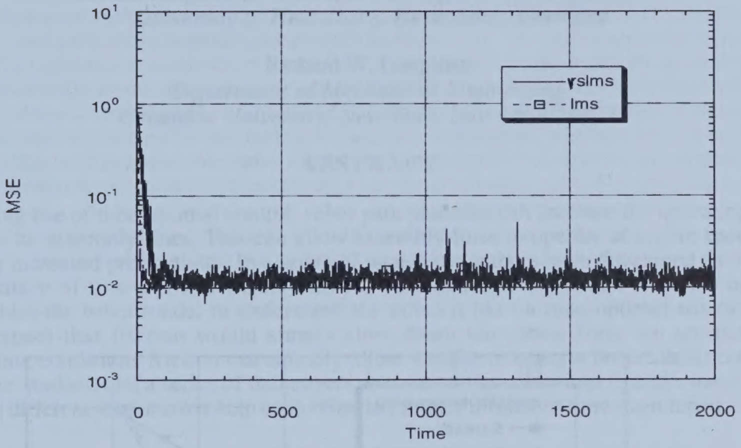


Figure 3: MSE learning curves for LMS with $\alpha = 0.33$ and VSLMS with $\underline{\alpha} = 0.08$ and $\overline{\alpha} = 0.33$.

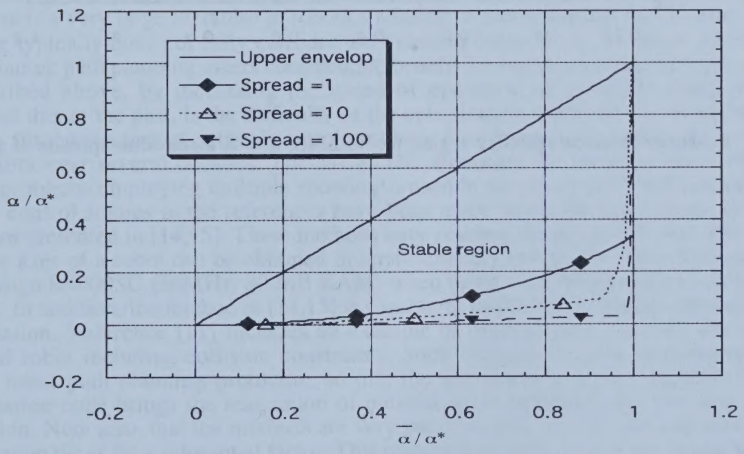


Figure 4: Prior stability stepsize region for VSLMS, $d = 2$, eigenvalue spreads of 1,10,100

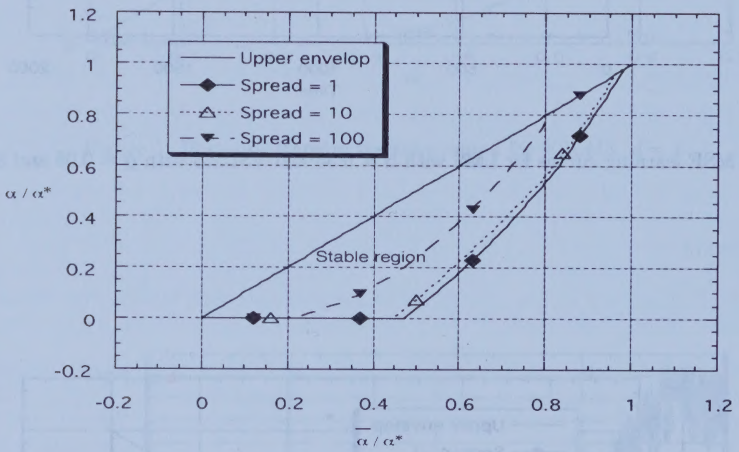


Figure 5: Posterior stepsize stability region for VSLMS, $d = 2$, eigenvalue spreads of 1,10,100

THE INFLUENCE OF FRICTION ON TIME OPTIMAL ROBOT TRAJECTORIES

Barbara Hartel, Marc Steinbach, and Hans Georg Bock
Interdisciplinary Center for Scientific Computing (IWR)
University of Heidelberg, Heidelberg, Germany

Richard W. Longman
Department of Mechanical Engineering
Columbia University, New York, New York 10027

ABSTRACT

By making use of time optimal control, robot path planning can increase the operating speeds of robots on assembly lines. This can allow assembly lines to operate at higher speeds with resulting increased productivity. In a series of papers the authors have developed an overview of the nature of time optimal trajectories for various kinds of robots. Here we introduce friction into the robot model to understand the effect it has on time optimal solutions. One might expect that friction would simply slow down the robot. Here we see that under appropriate conditions friction can actually allow a robot to operate faster. Polar coordinate robots are studied, and a series of maneuvers are treated that allow one to understand the way in which different joint motors help each other to produce minimum time trajectories.

INTRODUCTION

This paper is one in a series of studies by the authors begun in 1984, investigating the nature of time optimal robot trajectories, for polar robots, elbow robots, and scara robots (Refs. 1-8), and for satellite mounted robot path planning (Refs. 9-12). Friction is present in all mechanical systems, and all of the previous results in the references have ignored it. It is the purpose of this paper to study the effects of friction on time optimal robot trajectories. This study is one aspect of a substantial on-going research effort in optimal path planning, that aims to apply the methods to increasing the speed of operation of a specific robot operating in a press chain on the Mercedes Benz automobile manufacturing line. This robot is the slowest to accomplish its task, and hence, its speed determines the cycle time of the entire operation.

There is a very large literature in robotics relating to path planning. But to date, this path planning typically does not fully consider the dynamic behavior of the robot. Nevertheless, such dynamic path planning offers considerable benefit for increased productivity in industry, as described above, by increasing the speed of operation of assembly lines. What has prevented this in the past, is the difficulty of the optimization problems involved. Numerical methods for the solution of optimal control problems have been developed by the authors and co-workers over several decades. The MUSCOD algorithm for direct solution of optimal control problems employing multiple shooting is used in this study [13]. Various other robot optimal control studies in the references have been made using the more recent collocation algorithm presented in [14,15]. These methods have reached the point that time optimal paths for three axes of a robot can be obtained in approximately one minute on a work station, an IRIS Indigo R4000SC (50MHz, 80 MB RAM) when using 40 time intervals to discretize the control. In addition, the method in [14,15] is capable of including inequality constraints in the optimization. Reference [11] includes an example of optimal path planning for a satellite mounted robot including collision constraints. Such constraints are a fundamental part of typical robot path planning problems, so that the introduction of this capability into the optimization code brings the realization of optimal robot path planning one step nearer to realization. Note also, that the methods are very parallelizable, so that one can accelerate the computation times by a substantial factor. This opens the opportunity for use in real time robot operation.

THE ROBOT MODEL AND THE CHOICE OF MANEUVERS

The axes of a robot can usually be divided into those that are predominantly devoted to translational motion and those primarily devoted to attitude rotations. Since the attitude motion is usually fast compared to the translational motion, we concentrate on studying the time optimal control of the first three axes of a polar coordinate robot, as shown in Fig. 1.

It is difficult to obtain a general understanding of time optimal robot maneuvers because of the number of possible initial and final conditions that one might be interested in. Here, as in all of our previous works on time optimal robot problems, we decompose the problem in order to see the effects of the motors of one axis helping another axis accomplish its objective. Hence, we concentrate on maneuvers that would only require the use of one axis to connect the start and end points, and then we add another axis and study the way in which that axis can assist the first. Of course, in such a nonlinear dynamic system, superposition does not apply, but in a heuristic sense, any chosen robot maneuver can be thought of as a superposition of these different ways of interaction among the joints.

Three models are used. The first allows motion of the radial and base axes, the second allows motion of the radial and the elevation axes, and the third allows motion of the base and elevation angles. The equations for each of these models, together with the robot parameters are given in the appendix. The friction model is coulomb friction. The actuator limits are set to $|u_r| \leq 100\ N$, $|u_{\theta_1}| \leq 500\ Nm$, and $|u_{\theta_2}| \leq 500\ Nm$, unless otherwise noted.

OPTIMAL EXTENSION/RETRACTION WITH BASE MOTOR ASSISTING

The objective of the maneuvers studied in this section is simply to retract or extend the prismatic joint of the robot. If only the radial motor is used, then the time optimal trajectory is to pull in (push out) as hard as possible until half way to the endpoint, and then reverse the motor torque to push (pull) as hard as possible until arriving at the endpoint. The time optimal trajectory is different if you allow the base motor to assist the radial maneuver. This has been studied extensively in the absence of friction in [1,2] using direct methods, and using indirect methods which reveal the true switching structure. Here we study the influence of friction on these results.

Maneuver 1: The first maneuver retracts the arm in the horizontal plane from *r*=0.7 m (*r*=0.75 is fully extended), to *r*=0.3 m (*r*=0 puts the center of mass of the prismatic link over the axis of rotation of the base joint). The minimum times for this retraction maneuver are listed in Table 1 for various different levels of friction force/torque. One interesting conclusion from these numbers is that it is possible that the robot can actually perform a maneuver faster when it has friction to fight than when it does not. We will see other examples of this in the sequel. When the objective is to accelerate the link, friction hurts, and when the objective is to decelerate, it helps. It depends on relative sizes of this assistance and this detriment, whether introducing friction is of overall assistance.

Figure 2 shows the corresponding optimal trajectories for r and θ_1. The friction levels for the various curves are seen by reference to the Table. The entries on the table indicate the friction force or torque present when the joint is in motion. The base motor helps the radial motor by making a wiggle in the horizontal plane producing centrifugal force toward the end. This force enhances the robot's ability to decelerate the arm toward the end of the maneuver, and hence allows the radial motor to accelerate longer before having to start its braking. In the cases of R2, R3, and R4 it appears that friction delays the start of the wiggle until sometime well into the trajectory.

Figure 3 gives the associated optimal control histories for R0, R3 and R4. All of the numerical optimizations in this paper are done restricting the control variables to change their values only at 40 evenly spaced times during the maneuver. Since the controls cannot produce a bang at any point that is not the start of one of these 40 intervals, the optimal control for such a constrained control set need not be bang-bang even when the continuously variable control solution is. However, as shown in [2], generically in robot problems only one joint is totally bang-bang, the joint with the hardest task to accomplish. The other joint(s) have more time than needed to accomplish their task, and when this freedom can be used to help another axis, this joint too will be bang-bang. Otherwise, the control can go onto a singular arc. The true switching structure for R0 is given in [2]. There is a very short bang at the start, that begins the sideways rotation, and then the solution goes onto a singular arc until the time that creating centrifugal force can be of assistance. In addition to the characteristics introduced by the discretization in time of the admissible controls, the fact that the numerical methods used are iterative direct methods for which one adopts a stopping criterion, leaves open the possibility that solutions have not fully converged.

Examining Fig. 2, we see that introducing friction makes the amplitude of the wiggle smaller, and shifts its peak to later in the maneuver. This shift in the peak relates to the fact that there is rotational friction in the base joint which allows the wiggle motion to be stopped faster. And this means that under the right conditions the wiggle can be producing substantial centrifugal force later in the maneuver, when it is more helpful to the radial motor. And this can create the situation where the introduction of friction tallows a faster maneuver.

friction	radial joint	base joint	min. time retraction	min. time extension
R0	0 N	0 Nm	0.764 s	0.764 s
R1	20	100	0.7958	0.7573
R2	40	200	0.7568	0.7836
R3	60	300	0.7684	0.8618
R4	80	400	0.7911	1.091

Table 1. Minimum time as a function of friction level for the radial retraction and extension in the horizontal plane, maneuvers 1 and 2.

Maneuver 2: This is the reverse of maneuver 1, extending the arm from r=0.3 m to r=0.7 m. When there is no friction, one can show that the optimal retraction and extension trajectories must be the same with time reversed. But friction destroys this symmetry, influencing the two maneuvers in different ways, because friction always fights the motion, and the direction of motion is reversed on these trajectories.

The optimal trajectories for the radius versus time are given in Fig. 4, and the optimal base joint angle histories are in Fig. 5. This time the amplitude of the wiggle maneuver becomes larger rather than smaller as the friction level is increased. This time the centrifugal force is being used to help the arm produce the radial acceleration needed at the start of the trajectory. Friction makes it take longer to build up the amplitude of the wiggle, and the peaks are delayed in time. Since time is reversed by comparison to maneuver 1, this delay from friction moves the peaks in the opposite direction than in maneuver 1. The optimal maneuver times for this extension maneuver are also given in Table 1, and we see again that friction can decrease the time needed to make the extension. As seen in Fig. 6, for R1 the two peak angular velocities reached during the wiggle are larger than those produced without friction, and the associated increased centrifugal force allows a decreased total time. If the friction is increased to R2, the further delay in the time of these peaks together with the slight decrease in size has made the maneuver take longer than without friction.

Maneuver 3: This time we consider a retraction maneuver from r=0.7 to r=-0.2. This has the center of mass of the prismatic joint going past the axis of rotation of the base joint. After it passes this axis, producing centrifugal force is detrimental to the radial motor's task of braking the radial motion at the end. Hence, the rotational motor can only help between the times that the radial motor starts to brake, and the time that the center of mass reaches the axis of base rotation.

friction	radial joint	base joint	min. time retraction
R0	0 N	0 Nm	1.1993 s
R1	8	40	1.1807
R2	17	85	1.1643
R3	23	115	1.1504

Table 2. Minimum time as a function of friction level for the radial retraction past the origin in the horizontal plane, maneuver 3.

The minimum time solutions are given in Table 2, and we see that friction is helping substantially, and in the range investigated, the more friction, the faster the maneuver. Figures 7,8, and 9 show the optimal trajectories for the friction levels given in Table 2. Introducing a small friction level causes the rotational motion to stick at zero angle for some time at the start of the trajectory. The first wiggle amplitude is decreased by friction, but the second wiggle amplitude remains roughly the same and is delayed in time.

OPTIMAL EXTENSION/RETRACTION: ELEVATION MOTOR ASSISTING

Now consider retraction and extension maneuvers with the elevation motor allowed to assist the radial motion, rather than the base motor assisting. If there were no gravity, the optimal solutions for maneuvers 1, 2, and 3 would again be optimal in the vertical plane. However, gravity changes the behavior considerably. Without gravity, there are two locally optimal solutions, one with the wiggle maneuver going to the right, and the other with the wiggle maneuver going to the left. Reference [1] studies the effect of different gravity levels on these two solutions in the vertical plane without friction. As the gravity level is increased, the two solutions are modified and at a gravity level well below that of earth gravity the two solutions merge into one.

Maneuver 4: This maneuver uses the same start and end radii as maneuver 1. Figures 10 and 11 show the optimal trajectories for the radial motion and the elevation angle. The initial rise in elevation creates a situation where gravity helps the prismatic joint to accelerate inward. The dip below horizontal in the elevation angle later in the maneuver employs gravity to help stop the inward motion. In addition, the angular velocity produced toward the end creates centrifugal force that helps stop the motion. Friction delays both peaks, but the first peak can be larger without friction. For R3 the friction is so high that the elevation motor sticks at zero angle at the beginning of the maneuver, unable to fight both gravity torque and friction until the gravity torque has decreased as a result of the decreasing radius. As shown in Table 3, friction is detrimental to this retraction maneuver for all levels studied, presumably because friction in the elevation motor slows its ability to take advantage of gravity.

Maneuver 5: When the initial and final radii are reversed, as in maneuver 2, we get the optimal trajectories shown in Figs. 12 and 13. This time the initial objective of the elevation motor is to make the arm drop so that gravity will help pull the arm out. And this time gravity is helping the elevation motor overcome friction in its joint and produce the desired drop. The initial peak is delayed slightly, but it is made substantially larger, facilitated by the help from friction in stopping the elevation angle motion at the end of the maneuver. The fact that friction in the radial joint is helping to decelerate at the end seems to allow the elevation motor to put emphasis on coupling into gravity to help the acceleration portion of the trajectory. From Table 3 we see that friction decreases the time needed for this extension maneuver until we reach the R3 friction level. Note that extension maneuvers are faster than retraction maneuvers when allowing the elevation motor to assist, whereas the opposite was generally true in the case of the base motor assisting. Note also that the elevation motor is more effective in helping the radial motor than the base rotation motor is, producing lower minimum times, because it can make use of not only centrifugal and corriolis effects, but also gravity effects. Having studied separately how the θ_1 and the θ_2 motors can help in a radial maneuver, one can imagine the time optimal maneuver allowing both motors to assist, producing a compound wiggle in both angles [1].

friction	radial joint	base joint	min. time retraction	min. time extension
R0	0 N	0 Nm	0.7125 s	0.7125 s
R1	20	100	0.7425	0.7024
R2	40	200	0.8194	0.712
R3	49	245	0.8931	0.7255

Table 3. Minimum time as a function of friction level for the radial retraction and extension in the vertical plane, maneuvers 4 and 5.

OPTIMAL BASE ROTATION WITH THE RADIAL MOTOR ASSISTING

We now turn out attention to maneuvers whose start and end conditions would only require use of the base joint. For maneuver 6, the initial and final radial joint is at r=0.7, and the base joint must rotate by 50 degrees, and for maneuver 7 the radius is reduced to 0.1 m, and the base joint must rotate by 90 deg. The actuator limits for this section are reset to $|u_r| \le 500\ N$ and $|u_{\theta_1}| \le 100\ Nm$, in order to allow singular solutions for maneuver 7. Figure 14 shows the optimal radial motions for maneuver 6 for the friction levels indicated in Table 4, which also gives the minimum times needed to accomplish the rotation. The interpretation is that the radial motor pulls in, in order to decrease the inertia seen by the rotational motor. The corriolis forces produced by the radial motion also help in both the early acceleration and in the final deceleration. Figure 15 shows the corresponding motion for maneuver 7. This time the radial motor is sufficiently strong that it can cause the center of mass of the prismatic link to reach the base rotation axis. Going beyond this point starts to increase the inertia seen by the base motor, so that the radial motion goes onto a singular arc [1] with r=0. The solution without friction must be symmetric, and it overshoots zero because the quadratic effect on the inertia makes it best to get into the neighborhood of zero fast and allow an overshoot. The reverse is true on the return to the original radius. By going the wrong way first, and then getting a running start, one can stay in the neighborhood of zero longer. When friction is introduced, we do not expect that the inbound and outbound trajectories would look the same, and it appears that the "oveshoot" going out is larger than that going in, since going in has the help of friction to stick at the desired radius. Since the minimum is shallow, the numerical results do not give a perfect picture of the trajectory near zero radius. For R0 it does not stay at zero radius in the center although there should be such a portion, but it does when friction is introduced. The corresponding minimum maneuver times are given in Table 5.

friction	radial joint	base joint	time
R0	0 N	0 Nm	0.8739 s
R1	100	20	0.8928
R2	200	40	0.9575
R3	300	60	1.103
R4	400	80	1.4818

Table 4. Minimum times as a function of friction level for base rotation with the aid of the radial motor, nonsingular case, maneuver 6.

friction	radial joint	base joint	time
R0	0 N	0 Nm	1.0784 s
R1	217	34	1.147
R2	245	49	1.2374

Table 5. Minimum times as a function of friction level for base rotation with the aid of the radial motor, singular case, maneuver 7.

OPTIMAL BASE ROTATION WITH THE ELEVATION MOTOR ASSISTING

The elevation motor can also assist the base rotation motor, by decreasing the inertia it sees. This can be done by either lowering the arm (maneuver 8) or raising the arm (maneuver 9) with the θ_2 motor.

friction	base joint	elevation joint	lower solution	upper solution
R0	0 Nm	0 Nm	0.6152 s	0.631 s
R1	150	150	0.6494	0.6661
R2	250	250	0.7335	0.7425

Table 6. Minimum times as a function of friction level for the lower and upper solution for base rotation with the assistance of the elevation motor, maneuvers 8 and 9, respectively.

The corresponding optimal trajectories for θ_2 are given in Figs. 16 and 17, and the optimal times are given in Table 6, for the case of a 90 deg rotation in the base joint angle. Introducing friction prevents the elevation motor from reaching as big an angle, and therefore it cannot help as much, and the total times become longer. The upper solution (maneuver 9) is more strongly affected by friction since gravity is in the same direction as the friction when the elevation motor is attempting to raise the arm.

REFERENCES

1. Konzelmann, J., Bock, H.G. and Longman, R.W., "Time optimal trajectories of polar robot manipulators by direct methods," *Modeling and Simulation*, Instrument Society of America, 1989, 20, pp. 1933-1939.

2. Steinbach, M., Bock, H.G. and Longman, R.W., "Time-optimal extension or retraction in polar coordinate robots: a numerical analysis of the switching structure," *Journal of Optimization Theory and Applications*, March 1995. Also, *Proceedings of the AIAA Guidance, Navigation, and Control Conference*, Boston, 1989, pp. 883-894.

3. Konzelmann, J., Bock, H.G. and Longman, R.W., "Time optimal trajectories of elbow robots by direct methods," *Proceedings of the 1989 AIAA Guidance, Navigation, and Control Conference*, Boston, 1989, pp. 895-910.

4. Steinbach, M., Bock, H.G. and Longman, R.W., "Time optimal control of SCARA robots," *AIAA Guidance, Navigation and Control Conference, A Collection of Technical Papers*, Portland, Oregon, 1990, pp. 707-716.

5. B. Wolf (Hartel), *Numerische Bahnoptimierung von Robotern unter Berücksichtigung von Gelenkreibung*, Diplomarbeit, University of Heidelberg, 1995.

6. Mössner-Beigel, M., Steinbach, M, Bock, H. G., and Longman, R. W., "Time optimal path planning in polar robots with joint flexibility," AAS/AIAA Astryodynamics Conference, Sun Valley, Idaho, 1997, *Advances in the Astronautical Sciences*, to appear.

7. M. Mössner-Beigel, *Optimale steuerung für Industrieroboter unter Berücksichtigung der getriebebedingten Elastizität*, Diplomarbeit, University of Heidelberg, 1995.

8. Li, J., Longman, R. W., Steinbach, M., and Bock, H. G. "Implementing time optimal path planning with existing feedback controllers," *Proceedings of the 1997 AIAA Aerospace Sciences Meeting*, Reno, Nevada, January 1997, to appear.

9. Schulz, V.H., Bock, H.G. and Longman, R.W., "Optimal path planning for satellite mounted robot manipulators," *Advances in the Astronautical Sciences*, Vol. 82, 1993, pp. 311-329.

10. Schulz, V.H., Bock, H.G. and Longman, R.W., "Shortest paths for satellite mounted robots," *Computational Optimal Control*, ed R. Bulirsch & D. Kraft, International Series on Numerical Mathematics, Vol. 115, Birkhäuser Verlag, Basel, Boston, Berlin, 1994, pp. 357-366.

11. Longman, R. W., Schulz, V. H., and Bock, H. G., "Path Planning for Satellite Mounted Robots," Keynote Lecture, *Proceedings of the 3rd International Conference on Dynamics and Control of Structures in Space*, London, England, 27-31 May, 1996.

12. Schulz, V.H., "Optimal paths for satellite mounted robots," *ZAMM*, Special Issue on ICIAM 95, in press.

13. Bock, H.G. and Plitt, K.-J., *A Multiple Shooting Algorithm for Direct Solution of Optimal Control Problems*, Sonderforschungsbereich 72 Preprint, The University of Bonn, West Germany, 1983; also *Proceedings of the 9th IFAC World Congress*, Budapest, 1984.

14. Schulz, V., "A direct PRSQP method for path planning of satellite mounted robots," *ZAMM*, Special Issue on ICIAM 95, in press.

15. Schulz, V.H. *Reduced SQP Methods for Large-Scale Optimal Control Problems in DAE with Application to Path Planning Problems for Satellite Mounted Robots*, PhD Dissertation, University of Heidelberg, Heidelberg, Germany, 1995.

APPENDIX

The robot models for motion in the horizontal plane, vertical plane, and rotations only, are:

$$m_{LB}\ddot{r} - (m_{LB}r + m_L\ell)\,\dot{\theta}_1^2 = u_r + R_r$$
$$(I_3^1 + I_3^{23} + m_{LB}r^2 + 2m_L\ell r + m_L\ell^2)\ddot{\theta}_1 + 2(m_{LB}r + m_L\ell)\dot{r}\dot{\theta}_1 = u_{\theta_1} + R_{\theta_1} \tag{A1}$$

$$m_{LB}\ddot{r} - (m_{LB}r + m_L\ell)\,\dot{\theta}_1^2 = u_r + R_r - m_{LB}g\sin\theta_2$$
$$(I_1^{23} + m_{LB}r^2 + 2m_L\ell r + m_L\ell^2)\ddot{\theta}_2 + 2(m_{LB}r + m_L\ell)\dot{r}\dot{\theta}_2 = u_{\theta_2} + R_{\theta_2} - (m_{LB}r + m_L\ell)g\cos\theta_2 \tag{A2}$$

$$(I_3^1 + I_3^{23}\sin^2\theta_2 + (I_3^{23} + m_L\ell^2)\cos^2\theta_2)\ddot{\theta}_1 + 2(I_2^{23} - (I_3^{23} + m_L\ell))\cos\theta_2\sin\theta_2\dot{\theta}_1\dot{\theta}_2 = u_{\theta_1} + R_{\theta_1}$$
$$(I_1^{23} + m_L\ell^2)\ddot{\theta}_2 + (I_3^{23} - I_2^{23} + m_L\ell^2)\cos\theta_2\sin\theta_2\dot{\theta}_1^2 = u_{\theta_2} + R_{\theta_2} - m_L\ell g\cos\theta_2 \tag{A3}$$

where $I_1^{23} = I_1^2 + I_1^3$; $I_2^{23} = I_2^2 + I_2^3$; $m_{LB} = m_L + m_B$. Here the mass of the prismatic link is $m_B = 40\,kg$ and the load mass is $m_L = 0\,kg$, $\ell = 0.75$m is the half length of the prismatic link, g is the acceleration of gravity. The single superscripts 1, 2, 3 on the center of mass inertias represent the base, intermediate, and prismatic links respectively, and subscript 3 and 2 represents body fixed axes aligned with the vertical axis and along the prismatic link respectively, when θ_2 is zero in Fig. 1 (the third axis is completed in the right handed sense); $I_1^{23} = I_3^{23} = 18.5\text{kgm}^2$, $I_2^{23} = 0.12\text{kgm}^2$, $I_3^1 = 0\,kgm^2$. The R_r, R_{θ_1}, R_{θ_2} represent Coulomb friction.

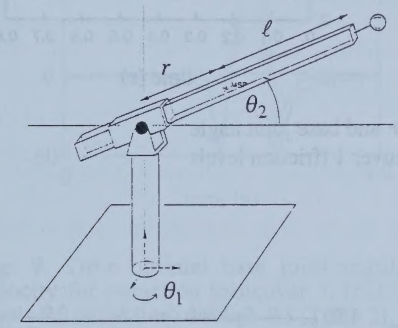


Fig. 1. Polar coordinate robot and notation.

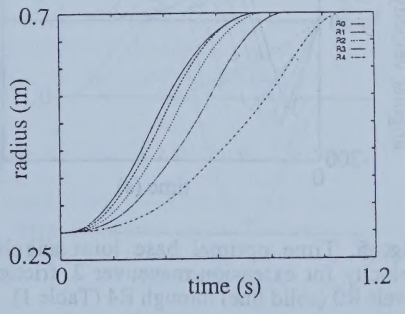


Fig. 4. Optimal radius r versus time for extension maneuver 2, friction levels R0 (solid line) through R4 (Table 1).

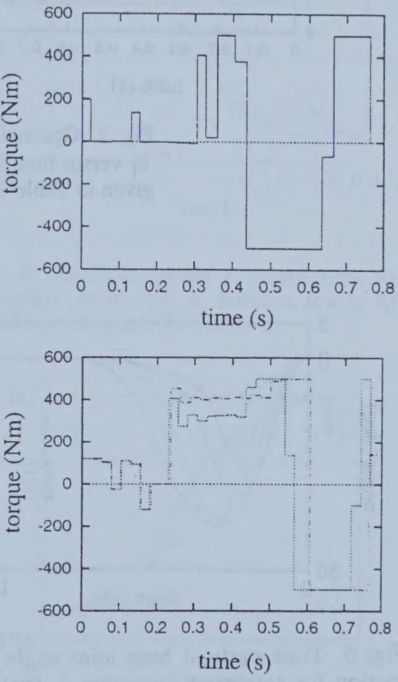


Fig. 3. Numerical solution for optimal base motor torque versus time for maneuver 1 for R0 (solid line), R3, and R4 (dot and dash line), key as in Fig. 2.

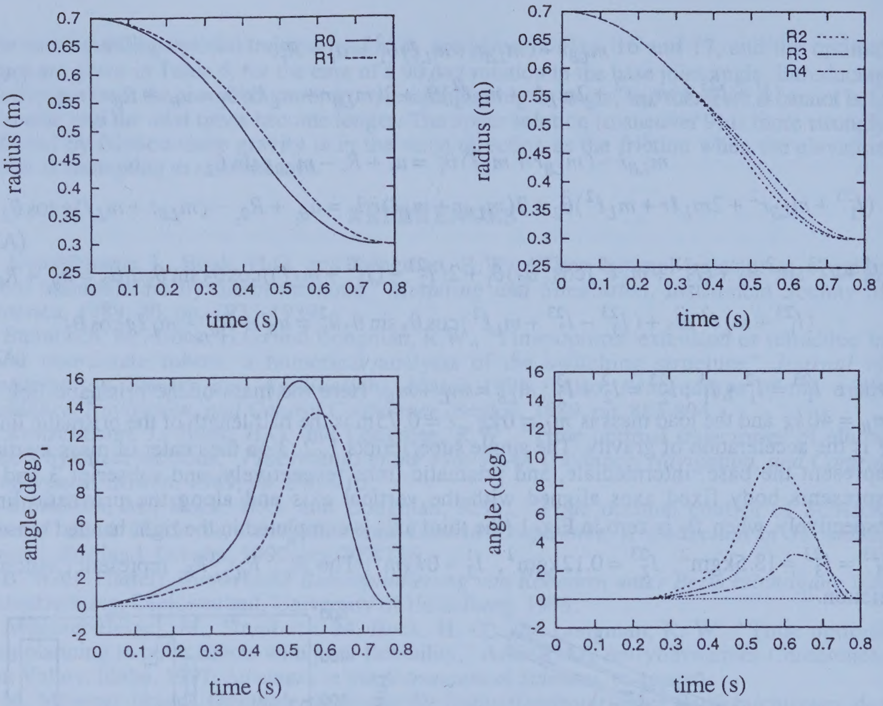
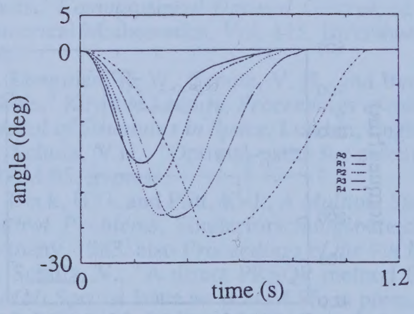


Fig. 2. Optimal radius r and base joint angle θ_1 versus time for maneuver 1 (friction levels given in Table 1).



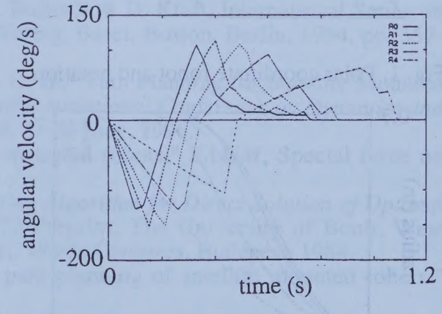


Fig. 5. Time optimal base joint angle θ_1 motion for extension maneuver 2, friction levels R0 (solid line) through R4 (Table 1).

Fig. 6. Time optimal base joint angular velocity for extension maneuver 2, friction levels R0 (solid line) through R4 (Table 1).

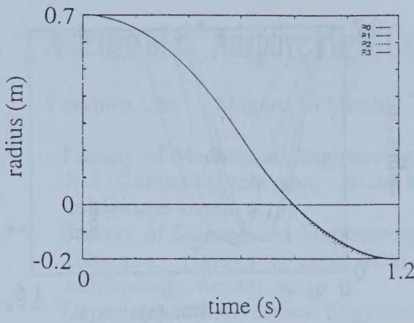


Fig. 7. Optimal radius r versus time for retraction maneuver 3, friction levels R0 (solid line) through R3 (Table 2).

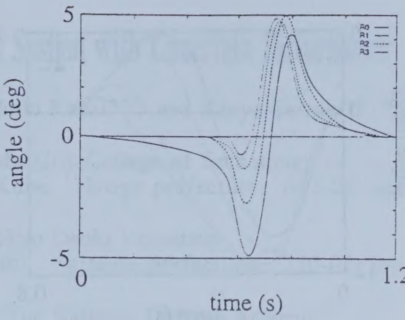


Fig. 8. Time optimal base joint angle θ_1 motion for retraction maneuver 3, friction levels R0 (solid line) through R3 (Table 2).

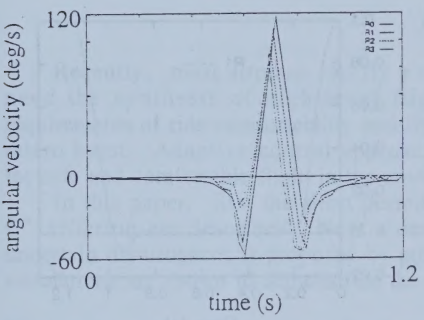


Fig. 9. Time optimal base joint angular velocity for retraction maneuver 3, friction levels R0 (solid line) through R3 (Table 2).

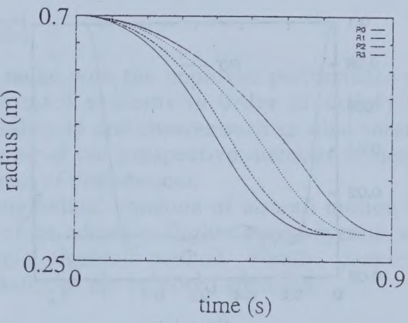


Fig. 10. Optimal radius r versus time for retraction maneuver 4, friction levels R0 (solid line) through R3 (Table 3).

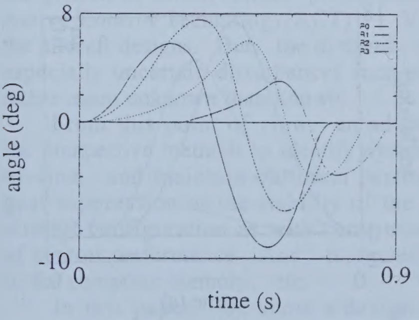


Fig. 11. Time optimal base joint angle θ_2 motion for retraction maneuver 4, friction levels R0 (solid line) through R3 (Table 3).

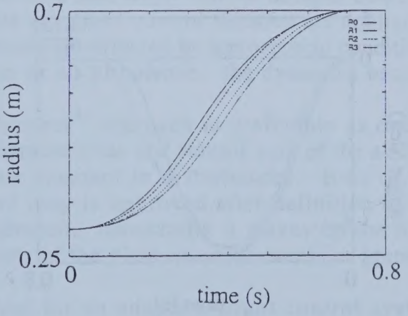


Fig. 12. Optimal radius r versus time for optimal extension maneuver 5, friction levels R0 (solid line) through R3 (Table 3).

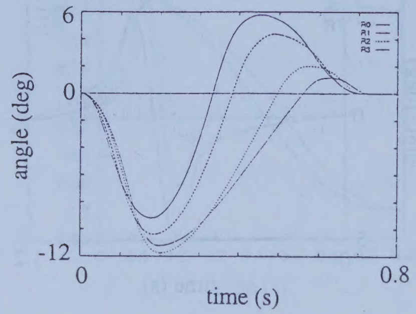


Fig. 13. Time optimal base joint angle θ_2 for extension maneuver 5, friction levels R0 (solid line) through R3 (Table 3).

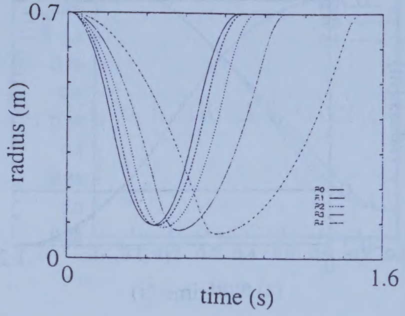


Fig. 14. Time optimal radial motion for 50 deg base rotation, maneuver 6, friction levels R0 (solid line) through R4 (Table 4).

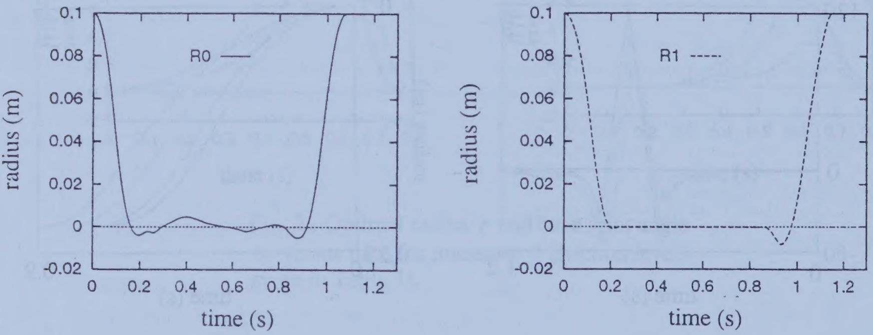


Fig. 15. The time optimal radial motion for the base rotation through 90 deg for friction levels R0 and R1, maneuver 7 (Table 5).

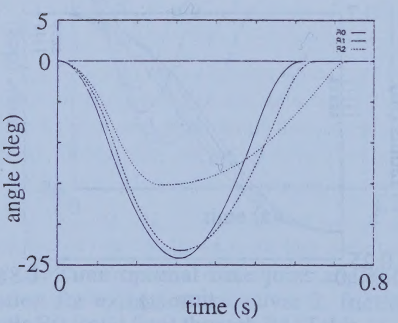


Fig. 16. The optimal elevation angle history for base rotation maneuver 8, friction levels R0 (solid line) through R2 (Table 6).

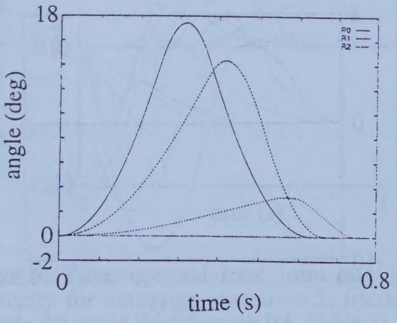


Fig. 17. The optimal elevation angle history for base rotation maneuver 9, friction levels R0 (solid line) through R2 (Table 6).

A design of C^* Adaptive Flight Control System with Uncertain Disturbance

Yasuhiro Osa*, Shigeru Uchikado**, Kimio Kanai*** and Kanya Tanaka****

* Faculty of Mechanical Engineering, Kobe City College of Technology
8-3 Gakuenhigashi-cho, Nishi-ku, Kobe, Hyogo prefecture, 651-21 Japan.
osa@kobe-kosen. ac. jp
** Faculty of Science and Engineering, Tokyo Denki University
Ishizaka, Hatoyama-cho, Hiki-county, Saitama prefecture, 350-03 Japan.
uchikado@j. dendai. ac. jp
*** Department of Aerospace Engineering, The National Defense Academy
1-10-20 Hashirimizu, Yokosuka, Kanagawa prefecture, 239 Japan.
kanai@cc. nda. ac. jp
**** Faculty of Engineering, Ehime University
3 Bunkyou-cho, Matsuyama, Ehime prefecture, 790-77 Japan.
eutanaka@dpc. ehime-u. ac. jp

Abstract

Recently, most airplane can fly a wide range with the improved performance and need the synthesis of high level flight control systems in order to satisfy the requirements of ride comfortability and flying safety in disturbances such as wind shear or micro burst. Adaptive control approach is one of the prospective methods of design for safe and comfortable flight in the environment of disturbances.

In this paper, first the short periodic longitudinal equation of aircraft motion and C^* criterion are described. Next a design of an adaptive flight control system with uncertain disturbances is proposed by polynomial algebraic method. Finally, numerical simulations are shown to demonstrate the feasibility of the proposed approach.

Keywords : Adaptive Control, Flight Control, Robust, Uncertainties

1. Introduction

Recently most airplanes are designed for high performance as a total system with integration of aerodynamic, propulsion, structural and flight control designs based on Active Control Technology(ACT)$^{1)}$. Variable advanced control theories are applied to the aircraft designs. But, the dynamics of aircraft are affected by atmospheric conditions especially uncertain disturbances such as gusts or air turbulence, the dynamics become rather more unknown than known.

From this point of view, an adaptive control$^{2)}$ approach is preferable as one of the prospective methods to identify dynamic characteristics and control gain of the aircraft on-line, and maintain optimum performance constant in disturbances. But, if the gust suppression or the stability of the closed loop is improved after definition of the aircraft configuration as with conventional aircraft, necessarily it makes up the limit of control performance. And, increased degrees of the plant cause an excessive increase in the computer memory, etc.

In this paper, we show a design method for an adaptive flight control system which needs increase of order of inner model$^{3)}$ and considers uncertain disturbance by polynomial algebraic method$^{4)}$.

In the second section, the short periodic longitudinal equation of aircraft motion is shown, and C^*-response$^{1)}$, which is expressed by linear addition of vertical acceleration at pilot position and pitch rate around the center of gravity, is considered as a controlled variable.

In the third section, problem formulation with uncertain disturbance is given. Next a design method of flight control system which compensates the stability of closed loop and alleviation of disturbance via polynomial algebraic method is presented and expanded to an adaptive system.

Finally the feasibility of the proposed method is demonstrated with numerical simulations under micro burst circumstances which cause aircraft accidents in landing or taking off.

2. Longitudinal Equation of Aircraft Motion[1)]

The short periodic longitudinal equation of aircraft motion can be expressed as follows :

$$
\begin{aligned}
\dot{w}(t) &= Z_w w(t) + U_o q(t) + Z_{\delta_e} \delta_e(t) \\
\dot{q}(t) &= (M_{\dot{w}} Z_w + M_w) w(t) + (M_{\dot{w}} U_o + M_q) q(t) + (M_{\dot{w}} Z_{\delta_e} + M_{\delta_e}) \delta_e(t)
\end{aligned}
\tag{1}
$$

where,

$w(t)$: vertical velocity [m/s], $q(t)$: pitch rate [rad/sec]

$\delta_e(t)$: elevator angle [rad], $u_e(t)$: input signal to actuator of elevator

g : gravity force [9.8 m/sec^2], U_o : air speed [m/sec]

and Z_w, Z_{δ_e}, M_w, $M_{\dot{w}}$, M_q, M_{δ_e} are longitudinal stability derivatives.

When the dynamics of the elevator actuator are considered as a first order system, the following equation can be obtained.

$$\dot{\delta}_e(t) = -(1/T) \cdot \delta_e(t) + (1/T) \cdot u_e(t) \tag{2}$$

where, T is time constant of the actuator.

Consider the state vector expression of Eq. (1) and Eq. (2). The state variables can selected as follows:

$$x^T(t) = [\, w(t),\ q(t),\ \delta_e(t) \,] \tag{3}$$

the state vector differential equation is expressed as follows :

$$\dot{x}(t) = Ax(t) + bu_e(t) \tag{4}$$

where,

$$A = \begin{bmatrix} Z_w & U_o & Z_{\delta_e} \\ M_{\dot{w}} Z_w + M_w & M_{\dot{w}} U_o + M_q & M_{\dot{w}} Z_{\delta_e} + M_{\delta_e} \\ 0 & 0 & -1/T \end{bmatrix}$$

$$b^T = [\, 0 \quad 0 \quad 1/T \,]$$

Next, consider the C^* criterion[1)] as the output of this system. The C^* criterion is one of criteria of flight performance. It considers the longitudinal short period mode of the airplane. It is defined as a criterion which evaluates the momentum of the airplane directly. Definition of the scalar which is expressed by linear addition of vertical acceleration at the pilot position and pitch rate around the center of gravity is called C^*-response. C^*-response for a step input is plotted with four ranges in Fig. 1.

$C^*(t)$ is expressed by the following equation with vertical acceleration at pilot position $n_{zp}(t)$ and pitch rate $q(t)$.

$$C^*(t) = n_{zp}(t) + (V_{co}/g) q(t) \tag{5}$$

Where V_{co} is cross-over velocity.

Also, with the relation between $n_{zp}(t)$ and $n_z(t)$:

$$n_{zp}(t) = n_z(t) + (l/g)\dot{q}(t) ,$$

Eq. (5) can be expressed by the following equation with normal vertical acceleration $n_z(t)$.

$$C^*(t) = n_z(t) + (V_{co}/g)q(t) + (l/g)\dot{q}(t) \quad (6)$$

Moreover, considering the relation :

$$n_z(t) = -w(t)/g + U_0\dot{q}(t)/g$$

and the state variables in Eq. (3) , $C^*(t)$ can be expressed as follows :

$$C^*(t) = c^T x(t) \quad (7)$$

where

$$c = \begin{bmatrix} [\,-Z_w + l(M_{\dot w}Z_w + M_w)\,]/g \\ [\,V_{co} + l(M_{\dot w}U_0 + M_q)\,]/g \\ [\,-Z_{\delta_e} + l(M_{\dot w}Z_{\delta_e} + M_{\delta_e})\,]/g \end{bmatrix}$$

and l is the distance between the center of gravity and pilot position.

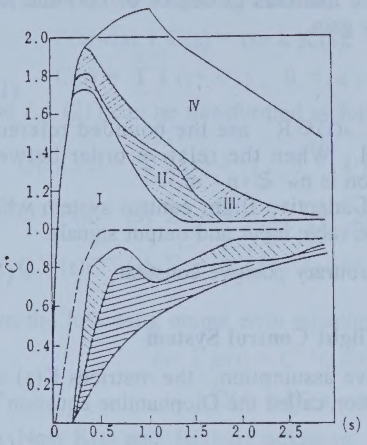


Fig. 1 C* criterion

Range I : Optimum response
Range II : Neither optimum nor danger
Range III : Not include Range I , II , IV
Range IV : Power approach response

3. Design of an Adaptive Flight Control System with Uncertain Disturbance

In this section, first problem formulation is given, a design method of flight control system with uncertain disturbance[5] by polynomial algebraic method is given, and the expansion of this system to adaptive control system is shown.

3. 1 Problem Formulation

Longitudinal equation of aircraft motion Eq. (4) and C^*-responce Eq. (7) can expressed with differential operator "s" as follows :

$$\begin{aligned} P(s)z(t) &= u_e(t) + [v(t) + \triangle v(t)] \\ C^*(t) &= R(s)z(t) \\ G(s) &= \frac{R(s)}{P(s)} = c^T(sI - A)^{-1}b \end{aligned} \quad (8)$$

where $u_e(t) \in R^1$ and $C^*(t) \in R^1$ are input and output signals, $z(t)$ is partial state. $v(t) \in R^1$ is deterministic disturbance[3] which satisfies $w(s)v(t) = 0$. The dynamics are described as follows :

$$w(s) = s^i + w_{i-1}s^{i-1} + \cdots + w_0 \quad (9)$$

It is assumed that the dynamics of uncertain part of disturbance are unknown and the upper limit of volume is known as follows :

$$\mathrm{Max}\,|\,\triangle v(t)\,| \leqq V < \infty, \quad V > 0 \quad (10)$$

Polynomials R(s) and P(s) are coprime, and are described as follows :

$$\begin{aligned} P(s) &= s^n + p_{n-1}s^{n-1} + \cdots + p_0 \\ R(s) &= r_m s^m + r_{m-1}s^{m-1} + \cdots + r_0 \end{aligned} \quad (11)$$

where, in the case of the system in section 2, the numbers of degree of P(s) and R(s) are $n = 3$ and $m = 2$. And r_m is the high frequency gain.

The reference model is selected as follows :

$$C^*_M(t) = G_M(s)u_M(t) \tag{12}$$

where $G_M(s) = R_M(s)P_M(s)^{-1}$, $u_M(t) \in R^1$ and $C^*_M(t) \in R^1$ are the bounded reference input and output. $P_M(s)$ is a stable polynomial. When the relative order between $R_M(s)$ and $P_M(s)$ is n_M, the model matching condition is $n_M \geqq n - m$.

The objective here is to develop a method of adaptive flight control system which satisfies the following equation, with only the observable input and output signals.

$$\lim_{t \to \infty} | C^*_M(t) - C^*(t) | = \delta \ : \ \text{arbitrary positive constant} \tag{13}$$

3. 2 Construction of Adaptive Flight Control System

As P(s) and R(s) are coprime due to the above assumption, the matrices K(s) and H(s) are uniquely determined by the following relation called the Diophantine equation[4),

$$\begin{aligned} &A(s)K(s)P(s) + A(s)H(s)R(s) \\ &\qquad = r_m Q(s)A(s)P(s) - A(s)Q(s)F(s)R(s) \end{aligned} \tag{14}$$

where, K(s), H(s) and Q(s) are the (n-1)th, (n-1)th and n-th order polynomials, F(s) is (n-m)th order stable polynomial. $A(s) = w(s)F(s) + G(s)$. w(s) is the inner model (the dynamics of deterministic disturbance) and A(s) set as G(s)/w(s) is proper and w(s)F(s) + G(s) becomes a stable polynomial.

Post-multiplying Eq. (14) by z(t), the following non-minimal expression is obtained.

$$\begin{aligned} &A(s)K(s)P(s)z(t) + A(s)H(s)R(s)z(t) \\ &\qquad = r_m Q(s)A(s)P(s)z(t) - A(s)Q(s)F(s)R(s)z(t) \end{aligned} \tag{15}$$

The orders of polynomial K(s), H(s) and Q(s) are the same as without disturbance. Also, the following relation can be obtained with Eq. (8) and $w(s)v(t) = 0$.

$$\begin{aligned} r_m u_e(t) &= K(s)Q(s)^{-1}u_e(t) + H(s)Q(s)^{-1}C^*(t) + F(s)C^*(t) \\ &\quad + [K(s) - r_m Q(s)]Q(s)^{-1}[v(t) + \triangle v(t)] \end{aligned} \tag{16}$$

Defining the output error $e(t) = C_M{}^*(t) - C^*(t)$, the following relation can be obtained.

$$\begin{aligned} F(s)e(t) &= -r_m u_e(t) + K(s)Q(s)^{-1}u_e(t) + H(s)Q(s)^{-1}C^*(t) \\ &\quad + F(s)C^*_M(t) + [K(s) - r_m Q(s)]Q(s)^{-1}[v(t) + \triangle v(t)] \end{aligned} \tag{17}$$

Now define the control input as

$$\begin{aligned} r_m u_e(t) &= K(s)Q(s)^{-1}u_e(t) + H(s)Q(s)^{-1}C^*(t) \\ &\quad + F(s)C_M{}^*(t) - G(s)w(s)^{-1}[C^*(t) - C_M{}^*(t)] + w(s)^{-1}C(s)\triangle\gamma(t) \end{aligned} \tag{18}$$

※ C(s) and $\triangle\gamma(t)$ are defined later

With this input, we have

$$[F(s)w(s) + G(s)]e(t) = [K(s) - r_m Q(s)]Q(s)^{-1}w(s)\triangle v(t) - C(s)\triangle\gamma(t) \tag{19}$$

Where $F(s)w(s) + G(s)$ is a stable polynomial. F(s) and G(s) are easily determined without the Heliminante matrix as follows. For example, a designer considers the characteristic equation of the closed loop, A(s) as follows :

$$A(s) = s^3 + f_2 s^2 + f_1 s + f_0$$

where if $w(s) = s^2$, F(s) and G(s) become as follows :

$$F(s) = s + f_2, \quad G(s) = f_1 s + f_0$$

And consider the following relation.

$$F(s)w(s) + G(s) = (s+\lambda)C(s),\quad \lambda > 0 \tag{20}$$

$$C(s) = \prod_{j=1}^{c-1}(s+\lambda'_j),\quad 0 < \lambda'_1 < \cdots < \lambda'_{c-1} \text{ and } c = n-m+i \tag{21}$$

Also Eq. (21) can be transformed as follows :

$$Q(s)C(s) = \prod_{j=1}^{q-1}(s+\lambda_j),\quad 0 < \lambda_1 < \cdots < \lambda_{q-1} \text{ and } q = 2n-m+i \tag{22}$$

$$Q^{-1}(s)C^{-1}(s)\sum_{i=0}^{q-1} s^i = 1 + \sum_{i=0}^{q-2}(s^i - c_i s^i)Q^{-1}(s)C^{-1}(s) \tag{23}$$

$$Q^{-1}(s)C^{-1}(s)s^i = \sum_{j=1}^{q-1} \bar{c}_i^j / (s+\lambda_j) \tag{24}$$

Then the following output error equation can be obtained.

$$(s+\lambda)e(t) = C^{-1}(s)D(s)\triangle v(t) - \triangle\gamma(t) \tag{25}$$

$$D(s) = [K(s) - r_m Q(s)]Q(s)^{-1}w(s)$$

Now K(s) and H(s) are unknown. The following control input which corresponds to Eq. (18) can be considered by using the adjustable parameters $\hat{K}(s, t)$, $\hat{H}(s, t)$, $\hat{r}_m(t)$.

$$\hat{r}_m(t)u_e(t) = \hat{K}(s, t)Q(s)^{-1}u_e(t) + \hat{H}(s, t)Q(s)^{-1}C^*(t) + F(s)C_M^*(t) + G(s)w(s)^{-1}e(t) + w(s)^{-1}C(s)\triangle\gamma(t) \tag{26}$$

In the same way as Eq. (25), the following output error equation can be obtained.

$$(s+\lambda)e(t) = w(s)C^{-1}(s)\phi^T(t)\xi(t) + C^{-1}(s)D(s)\triangle v(t) - \triangle\gamma(t) \tag{27}$$

where

$$\phi(t) = \theta - \hat{\theta}(t)$$

$$\theta^T = (r_m, k_0, k_1, \cdots, k_{n-1}, h_0, h_1, \cdots, h_{n-1})$$

$$\hat{\theta}(t)^T = [\hat{r}_m(t), \hat{k}_0(t), \hat{k}_1(t), \cdots, \hat{k}_{n-1}(t), \hat{h}_0(t), \hat{h}_1(t), \cdots, \hat{h}_{n-1}(t)]$$

$$\xi(t)^T = [-u_e(t), u_{e2}(t), su_{e2}(t), \cdots, s^{n-1}u_{e2}(t), C^*_3(t), sC^*_3(t), \cdots, s^{n-1}C^*_3(t)]$$

$$u_{e2}(t) = Q(s)^{-1}u_e(t)$$

$$C^*_3(t) = Q(s)^{-1}C^*(t)$$

The adjustable parameters are estimated with the following method. First, define the augmented error equation as follows :

$$(s+\lambda)\varepsilon(t) = (s+\lambda)e(t) + \{\phi^T(t)w(s)C^{-1}(s) - w(s)C^{-1}(s)\phi^T(t)\}\xi(t) \tag{28}$$

and

$$(s+\lambda)\varepsilon(t) = \phi^T(t)\zeta(t) + C^{-1}(s)D(s)\triangle v(t) - \triangle\gamma(t) \tag{29}$$

where $\zeta(t) = w(s)C^{-1}(s)\xi(t)$.

Next, define the adaptive law[5)] as follows :

$$\dot{\phi}(t) = -\Gamma\varepsilon(t)\zeta(t),\quad \Gamma = \Gamma^T > 0 \tag{30}$$

The Liapnov function are selected as follows :

$$L(t) = (1/2)[\varepsilon(t)^2 + \phi(t)^T\Gamma^{-1}\phi(t)] \tag{31}$$

Also the time differential of the above function becomes as follows :

$$\dot{L}(t) = -\lambda\varepsilon(t)^2 + \varepsilon(t)[C^{-1}(s)D(s)\triangle v(t) - \triangle\gamma(t)] \tag{32}$$

Considering the following state variable filter[5],

$$\dot{V}_v(t) = -\lambda_1 V_v(t) + V \tag{33}$$

the following relation can be obtained.

$$C^{-1}(s)D(s)\triangle v(t) \leqq (n+1)M[\sum_{i=0}^{q-2}\sum_{j=1}^{q-1} |\bar{c}_i^j| V_v(t)] \tag{34}$$

where M is the maximum value of coefficients in the polynomial $[K(s) - r_m Q(s)]$. Therefore if $\triangle\gamma(t)$ is developed as follows :

$$\triangle\gamma(t) = M_T(t)\ \{(n+1)[\sum_{i=0}^{q-2}\sum_{j=1}^{q-1} |\bar{c}_i^j| V_v(t)]\}, \tag{35}$$

where

$$M_T(t) = \begin{cases} M & \varepsilon(t) > \delta \\ 0 & |\varepsilon(t)| \leqq \delta \\ -M & \varepsilon(t) < \delta \end{cases}$$

the following equation which means the Liapnov function can be obtained.

$$\dot{L}(t) \leqq 0 \tag{36}$$

Hence, when $\varepsilon(t) < \delta$, $\dot{\phi}(t)$ converges to 0 in finite time, the augmented error $\varepsilon(t)$ becomes equal to e(t) with Eq. (28), the following control objective can be obtained.

$$|e(t)| < \delta \tag{37}$$

4. Application to a Small, High Speed Aircraft

In this paragraph, we attempt to apply the proposed method to the flight control system of a small, high speed aircraft, and investigate the feasibility of the proposed approach.

First, considering flight condition[1] I [altitude : 6000 m, velocity : Mach 0.8, V_{co} : 99.8 m/s, T : 0.05 1/sec, α_{trim} : 2.6°], the plant parameters of aircraft are given as follows :

[Flight Condition I]

$$A = \begin{bmatrix} -0.882 & 254.4 & -47.394 \\ -0.008 & -1.217 & -22.062 \\ 0 & 0 & -20 \end{bmatrix}$$

$$b^T = [\ 0 \quad 0 \quad 20\ ]$$

$$c^T = [\ 0.085 \quad 9.486 \quad -7.966\ ]$$

The reference model is selected as the following 2nd order system having damping ratio 0.9 and undamped natural frequency 5.2 rad/s.

$$G_M(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n + \omega_n^2} \tag{38}$$

The reference input is given the unit step function.

Micro burst which causes aircraft accidents in landing or taking off can be considered the deterministic (step) disturbance in Fig. 2. In the beginning headwind, next headwind and wind shear, wind shear and finally fair wind and wind shear are given. The elimination of deterministic (step) disturbance is satisfied by giving w(s) = s. Also the uncertain element of disturbance, $\triangle v(t) = \sin 10t$ or $2\sin 10t$, is added to the above disturbance.

Simulation results are shown in Fig. 3～7. Fig. 4 and Fig. 6 show that the effect of deterministic disturbance is eliminated, the oscillation by uncertain disturbance is alleviated within a small range and the control objectctive of this study is achieved.

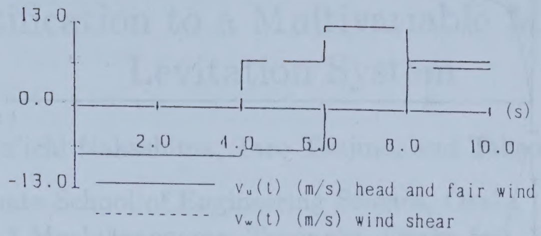


Fig. 2 Deterministic disturbance (Micro burst)

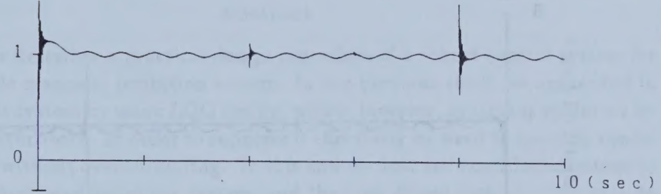


Fig. 3 C^*- response without compensator for uncertain disturbance $\triangle v(t) = \sin 10t$

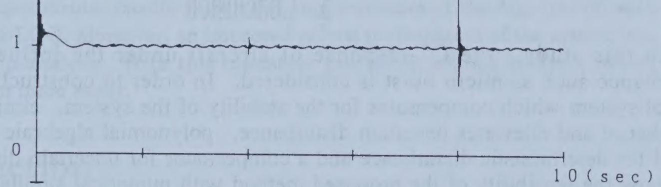


Fig. 4 C^*- response with compensator for uncertain disturbance $\triangle v(t) = \sin 10t$

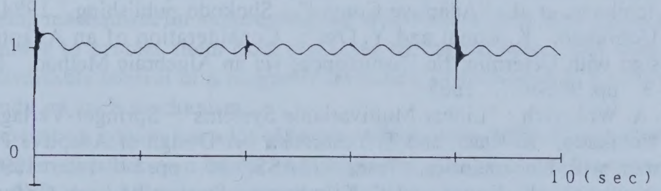


Fig. 5 C^*- response without compensator for uncertain disturbance $\triangle v(t) = 2\sin 10t$

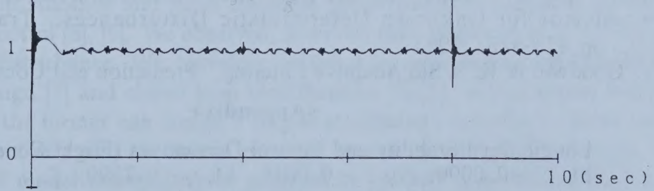
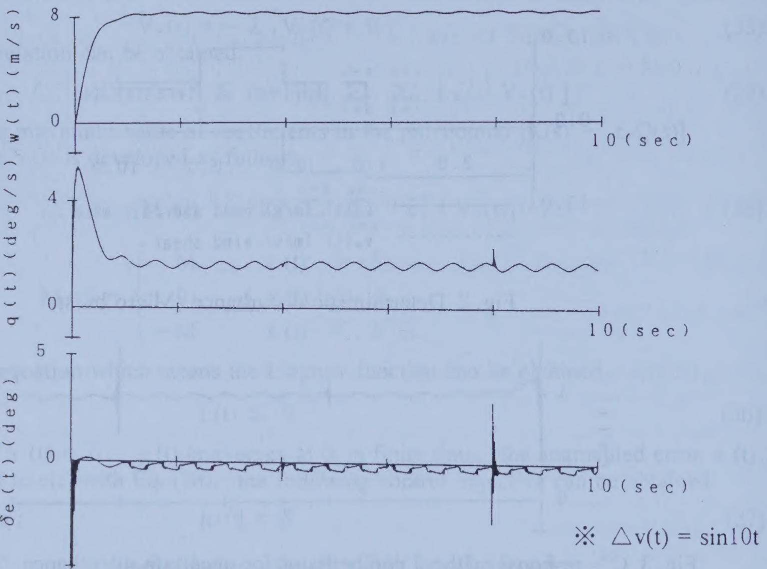


Fig. 6 C^*- response with compensator for uncertain disturbance $\triangle v(t) = 2\sin 10t$



※ $\triangle v(t) = \sin 10t$

Fig. 7 The response of state variables with compensator for uncertain disturbance

5. Conclusion

In this study, the C^*-response of aircraft under the influence of uncertain disturbance such as micro burst is considered. In order to construct an adaptive flight control system which compensates for the stability of the system, elminates deterministic disturbance and alleviates uncertain disturbance, polynomial algebraic method with inner model for deterministic disturbance and a compensator for uncertain disturbance are used. We show the feasibility of the proposed method with numerical simulations.

<References>

1) K. Kanai : "Flight Control", Maki publishing, 1985
2) K. Ichikawa et al : "Adaptive Control", Shokodo publishing, 1984
3) S. Uchikado, K. Kanai and Y. Osa : Consideration of an Adaptive Control System Design with Deterministic Disturbances via an Algebraic Method, Trans. SICE Japan, 24-9, pp. 985-987, 1988
4) W. A. Wolovich : "Linear Multivariable Systems", Springer-Verlag, 1974
5) S. Uchikado, K. Kanai and T. Yoshiduka : A Design of Adaptive Flight Control System with Uncertainties, Trans. JSASS, 40, pp. 143-149, 1992
6) T. Noguchi, K. Kanai and S. Kikukawa : Design of Low Order Adaptive Control System with Relay Type Adaptive Law and Application to Flight Control, Trans. SICE Japan, 22, pp. 351-357, 1986
7) H. Ohmori and A. Sano : Model Reference Adaptive Control System Including Compensator for Unknown Deterministic Disturbances, Trans. SICE Japan, 20-10, pp. 919-925, 1984
8) G. C. Goodwin & K. S. Sin;Adaptive Filtering, Prediction and Control", 1984

<Appendix>

Longitudinal Stability and Control Derivatives (Flight Condition I)

M_w : -0.0096, $M_{\dot{w}}$: -0.0018, M_q : -0.7590, Z_w : -0.8820

M_{δ_e} : -22.1470, Z_{δ_e} : -47.394, l : 5.7 m, g : 9.8 m/sec^2

Application of $\mathcal{H}_\infty$ Control and Closed Loop Identification to a Multivariable Magnetic Levitation System

Ken'ichi Nakashima, Taro Tsujino, and Takao Fujii

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560, Japan
E-mail: fujii@sys.es.osaka-u.ac.jp

Abstract

This paper describes a practical design procedure of a robust control system for a multivariable magnetic levitation system. In our previous study we succeeded in stabilizing this system by using LQG design, which, however, generates spillovers by an impulse disturbance. In order to suppress it effectively we need to quantify model uncertainties without overestimating. To this end we first use exact linearization to compensate nonlinearities of the system, and then use closed loop identification to estimate vibration characteristics of the system as ignored in its physical modeling. As a result, we can easily design an $\mathcal{H}_\infty$ servo controller in a two degree of freedom control configuration which ensures both robust stability and improved transient responses. Experimental results show good step responses of the $\mathcal{H}_\infty$ control such as those of the LQG. Moreover, an improved robust performance of the system was confirmed by μ-analysis. With this consistent procedure, we can realize a robust control system more easily.

1 Introduction

Magnetic levitation systems without mechanical contacts have been studied recently in various industries [1], [2]. In steel plants for example, a magnetic levitation technique is studied as a promising mechanism for suppressing the vibrations of thin steel plates under transportation in order to protect them from scratches on the surfaces. This motivates us to study a multivariable control of a magnetic levitation system with a Y shape iron plate as a basic study on such mechanism.

The magnetic levitation system has a lot of factors which cause difficulties in stabilization: 1) Mutual interferences between inputs and outputs of the system, 2) Nonlinearities of magnetic forces, 3) A destabilizing phenomenon of the close loop system as caused by spillovers. In our previous study [3], we used exact linearization [4] to compensate the nonlinearities of 2) and obtained a linearized model which is diagonalized against 1). In fact, its closed loop stability and good transient properties were achieved by using an ILQ-based LQG control [5], [6]. We observed, however, that spillovers of 3) were generated by an impulse disturbance. We, therefore, solved it by applying a combination of $\mathcal{H}_\infty$ control system design [7] and closed loop identification [8]-[11] in our recent study [12]. As is well-known, the former can design a robust stabilizing controller so as to suppress the vibrations effectively and the latter provides necessary information for the design by quantifying the model uncertainty in addition to identify a model for the unstable system. This paper aims to construct a practical and consistent procedure covering from modeling to design for robust control systems based on the experimental knowledge.

We have already examined the closed loop identification experimentally from the viewpoint of a robust control system design based on identifiability for the closed loop identification [9] by L. Ljung *et al.*. To our knowledge, no reports have been published so far on this identification for unstable mechanical systems uncertainties without overestimating in some limited frequency range. Based on this information, we then apply a mixed sensitivity problem of $\mathcal{H}_\infty$ control in the configuration of a two degree of freedom control servosystem [13]-[15]. In addition we analyze the robust performance of the improved control system by μ-analysis, and it turns out that we can design a superior $\mathcal{H}_\infty$ controller which ensures both robust stability against model uncertainty and robust performance with respect to tracking of transient responses to desired responses.

This paper is organized as follows. A control mechanism of our system are explained in section 2. First, the exact linearized model is obtained by compensating nonlinearities of the system, and second the identified model for the characteristics of natural vibrations is obtained by closed loop identification. Last, the relative difference in gain between these models is regarded as multiplicative uncertainties of the linearized model in section 3. We apply the mixed sensitivity problem of $\mathcal{H}_\infty$ control so as to ensure robust stability of the system, and hence can expect that the resulting closed loop system suppresses the spillovers. In addition, we design a two degree of freedom servosystem which can specify both the characteristics of disturbance reductions and the transient responses independently due to improved output responses in section 4. The μ-analysis results guarantee robust performance of the improved control system. Experimental results are presented in section 5. Finally, the conclusion and the future study is described in section 6.

2 Experimental apparatus

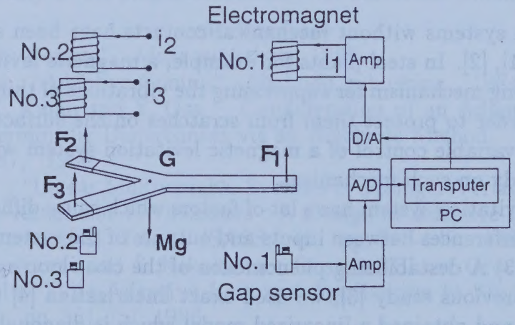


Figure 1: Schematic diagram of magnetic levitation system

Fig. 1 depicts a schematic diagram of our magnetic levitation system with a Y shape iron plate as a levitated vehicle. The control objective is to levitate the plate at various reference position and angles which may vary in a step wise, or in a rough term, the attitude control of the plate under levitation. The levitated vehicle, the Y shape iron plate as shown in Fig. 2, is made of aluminum with small pieces of iron plate mounted at the edges. The total mass M of the plate is about 2 kg.

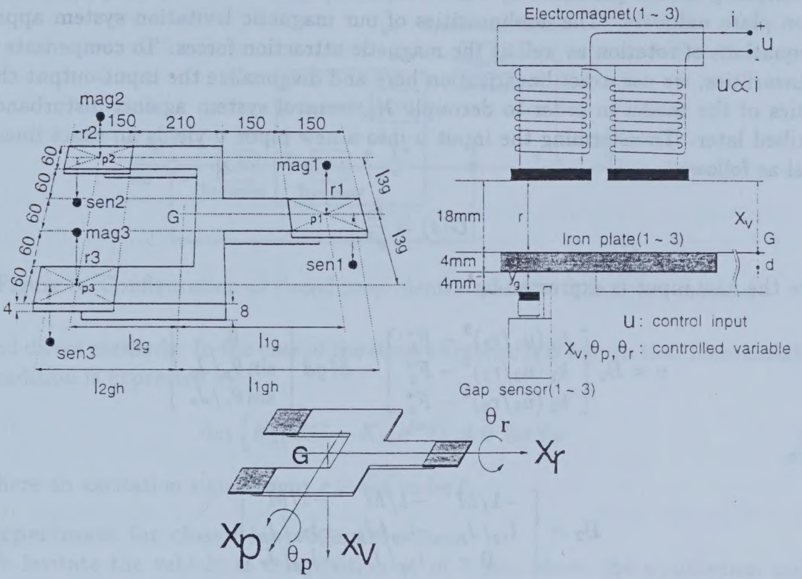

Electromagnet(1 ~ 3)
i
+
U
-
U∝i
mag2
150
210
150
150
r2
p2
60
60
60
60
60
60
mag1
r1
sen2
G
p1
39
39
mag3
r3
sen1
p3
4
8
sen3
l2g
l1g
l2gh
l1gh
18mm
r
Xv
Iron plate(1 ~ 3)
G
4mm
d
4mm
yg
U : control input
Xv, θp, θr : controlled variable
Gap sensor(1 ~ 3)
θr
G
Xr
Xp
θp
XV

Figure 2: Levitated vehicle and coordinate axis

3 Modeling

In this section we first apply exact linearization to derive a linear nominal model based on physical laws and then apply closed loop identification to identify the model uncertainty corresponding to vibration modes of the vehicle.

3.1 Modeling of Rigid Mode with Exact Linearization

A neglected part of nonlinearities associated with linearization using Taylor's series expansion would generally cause a degradation of the resulting performance because an effective working range of this linearization is limited around an equilibrium point. In the case of our system, exact linearization using an input transformation alone can compensate it by minimizing the modeling error as produced by the approximate linearization.

We define state vector x, input vector u and output vector y as follows:

$$x = \begin{bmatrix} x_v - x_v^* \\ \theta_p \\ \theta_r \\ \dot{x}_v \\ \dot{\theta}_p \\ \dot{\theta}_r \end{bmatrix} = \begin{bmatrix} y \\ \dot{y} \end{bmatrix}, u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \tag{1}$$

where x_v is a vertical gap length between the electromagnet and the plate at the point right above the center of gravity, while θ_p and θ_r are the pitching and rotating angles (see Fig. 2 for the notations) and x_v^* is the vertical gap length at an equilibrium state.

The output y here is produced by three variables which determine the positions of levitation plate uniquely. The nonlinearities of our magnetic levitation system appear in the equations of rotation as well as the magnetic attraction forces. To compensate these nonlinearlities, we use exact linearization here and diagonalize the input-output characteristics of the model in order to decouple $\mathcal{H}_\infty$ control system against disturbances as described later. Transforming the input u into a new input v yields an exact linearized model as follows:

$$G(s) = \frac{I_3}{s^2} \tag{2}$$

where the new input is expressed by

$$v = B_2 \begin{bmatrix} k_1 (u_1/r_1)^2 - F_1^* \\ k_2 (u_2/r_2)^2 - F_2^* \\ k_3 (u_3/r_3)^2 - F_3^* \end{bmatrix} - Mgd \begin{bmatrix} 0 \\ \sin\theta_p/J_p \\ \sin\theta_r/J_r \end{bmatrix} \tag{3}$$

where

$$B_2 = \begin{bmatrix} -1/M & -1/M & -1/M \\ l_{1g}/J_p & -l_{2g}/J_p & -l_{2g}/J_p \\ 0 & l_{3g}/J_r & -l_{3g}/J_r \end{bmatrix}. \tag{4}$$

Note that the model (2) has no such modeling errors as produced by linear approximation, and in addition the model parameters do not vary depending on the location of the levitated vehicle. Hence, we may expect that the performance of the control system would be more effective than that based on the linear approximation model.

3.2 Quantification of Flexible Mode with Closed Loop Identification

In our previous study we applied LQG control with ILQ state feedback to our system, and observed by signal processing of stationary output data that the closed loop system has natural vibrations of lateral and longitudinal directions nearby 64 and 74 Hz respectively, as well as sensor noises at 60 Hz of source frequency. If an impulse disturbance is added to the system, the system becomes unstable gradually; these so-called spillovers are generated due to the neglected vibration modes. This is why we applied closed loop identification [8]-[11] to obtain an uncertainty model for robust stabilization systematically. We treat the indirect method which can easily utilize the above a-priori information of our previous study [12]. The method can indirectly obtain a true plant $\tilde{G}(s)$ by identifying two time series models made from stationary output data $\tilde{y}$ in a position where the largest vibration occurs, provided reference inputs for the identification is not used so as not to cause spillovers. The details for the case of direct method are omitted here.

The procedure for obtaining the uncertainty model $\Delta_T(s)$ with indirect method in Fig. 3 is as follows:

1) **Identifiability**
We prepare two different LQG controllers $K_1(s)$, $K_2(s)$ based on ILQ design method which satisfy identifiability for closed loop identification [9] in both cases of indirect

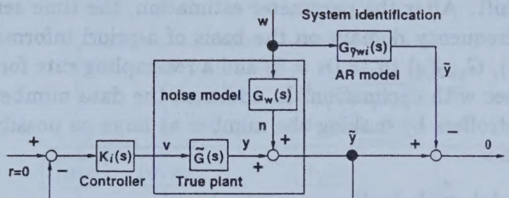


Figure 3: Configuration of closed loop identification with indirect method

and direct methods. In the case of the square system, it is known that identifiability condition is expressed by

$$\det\left\{K_2(e^{j\omega}) - K_1(e^{j\omega})\right\} \neq 0 \ \text{for} \ \forall\omega \tag{5}$$

where an excitation signal input r is set to be 0.

2) **Experiment for closed loop identification**
We levitate the vehicle at the position y_1 of 3 mm above the equilibrium point x_v^* nearby electromagnets where natural vibrations are large, and then switch a controller to another under stable operation to obtain the corresponding output data.

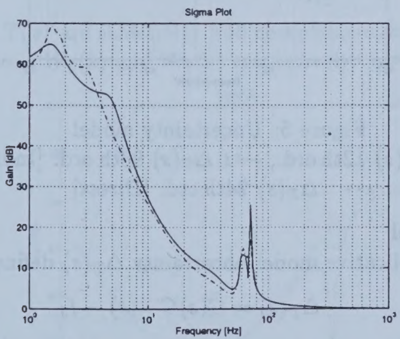


Figure 4: Frequency characteristics of time series models
— : $G_{\tilde{y}w1}(s)$ 18th ord., - · - : $G_{\tilde{y}w2}(s)$ 18th ord.

3a) **Identification of time series model**
We identify the time series models $G_{\tilde{y}w1}(s)$, $G_{\tilde{y}w2}(s)$ for each controller as multivariable Auto-Regressive (AR) models in the case of 3 inputs/outputs where w is assumed a white noise (see Fig.4 for their frequency characteristics and compare them with the above a-priori information). This is used with the Least Squares Method (LSM) as a way to identify the parameters:

$$A(q)\tilde{y}(t) = w(t) \tag{6}$$

where $A(q) = I_3 + A_1 q^{-1} + \cdots + A_n q^{-n}$; q^{-1} is a shift operator and n is the degree of the shift. After the parameter estimation, the time series model $G_{\tilde{y}w}(s)$ is obtained in frequency domain on the basis of a-priori information. We set the order of $G_{yw1}(s)$, $G_{yw2}(s)$ to 18 ($n = 6$) and a resampling rate for the identification to 1.024 × 5 msec with decimation. In addition, the data number is selected 1.5×10^4 in each controllers by making the number as large as possible to improve the signal/noise ratio.

3b) **Identified model with Indirect method**

A true plant $\tilde{G}(s)$ is indirectly determined by the following equation, which is derived from the fact that a noise model $G_w(s)$ does not change after switching a controller to another.

$$\begin{aligned} G_w(s) &= [I + \tilde{G}(s)K_1(s)]G_{\tilde{y}w1}(s) & (7)\\ &= [I + \tilde{G}(s)K_2(s)]G_{\tilde{y}w2}(s) & (8)\\ \Rightarrow \tilde{G}(s) &= [G_{yw2}(s) - G_{yw1}(s)] \times \\ &\quad [K_1(s)G_{yw1}(s) - K_2(s)G_{yw2}(s)]^{-1} & (9) \end{aligned}$$

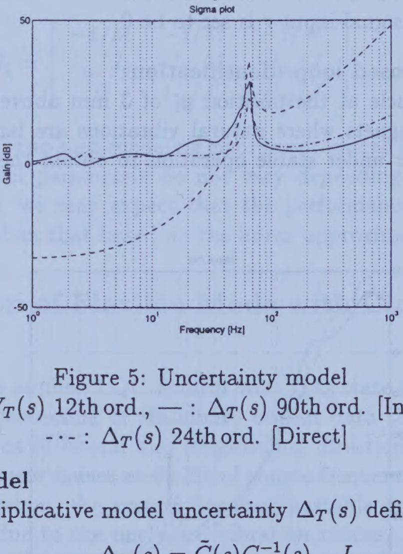


Figure 5: Uncertainty model
- - - : $W_T(s)$ 12th ord., — : $\Delta_T(s)$ 90th ord. [Indirect],
- · - : $\Delta_T(s)$ 24th ord. [Direct]

4) **Uncertainty model**

We identify a multiplicative model uncertainty $\Delta_T(s)$ defined by

$$\Delta_T(s) = \tilde{G}(s)G^{-1}(s) - I_3 \qquad (10)$$

The characteristics of our identified uncertainty model is shown in Fig. 5 for both methods. The results may be summarized as follows. First, it is known empirically from sampling time T_s that a frequency band of high accuracy identification lies from $1/100T_s$ to $1/5T_s$ of about 1 decade, *i.e.* from 10 to 200 Hz. Although the accuracy of identification at low frequencies is not enough due to a systematic error which does not depend on the data number, it is not such a serious problem; the weighting function for robust control system design should be estimated so as to have low gains at these frequencies since its behavior is governed by the rigid-body motion. Second, when using the direct method it was difficult to obtain a preferable order of the numerator of the identified ARX model. In the case of identifying an unstable system, the indirect method is superior to the direct method, in that it enables us to utilize a lot of a-priori information and estimate the uncertainty with little effort.

4 Robust control system design

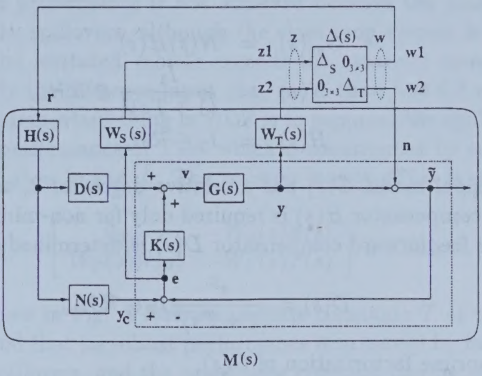


Figure 6: Robust performance analysis framework

The procedure for designing a two degree of freedom $\mathcal{H}_\infty$ servo control system shown in Fig. 5 is as follows:

Step 1) **Characteristics of noise and disturbance rejection $G_{yn}(s)$**

We consider a mixed sensitivity problem of $\mathcal{H}_\infty$ control as follows:

$$\left\| \begin{array}{c} \gamma_{opt} W_S(s) S(s) \\ W_T(s) T(s) \end{array} \right\|_\infty < 1 \qquad (11)$$

where $S(s)$ and $T(s)$ are sensitivity and complementary sensitivity functions with weighting functions $W_S(s)$ and $W_T(s)$ respectively.

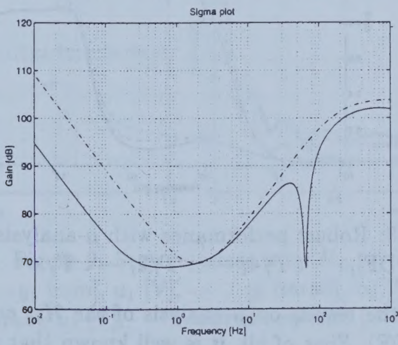


Figure 7: The gain plots of $K(s)$ with $\mathcal{H}_\infty$

— : $\gamma_{opt} = 1.01$ [$\mathcal{H}_\infty$] 6th ord., - · - : [ILQ] 6th ord.

An $\mathcal{H}_\infty$ servo controller $K(s)$ is determined by solving γ-iteration for (11) while tuning the weighting function $W_S(s)$ under the fixed weighting function $W_T(s)$. The $\mathcal{H}_\infty$ servo controller has the same degree 21 as the sum of the degrees of $W_T(s)G(s)$, $W_S(s)$ and its internal model $\alpha(s)$, and it plays a role of a notch filter with notches at frequencies where there exists natural vibrations as shown in Fig. 7. (See [12] for the details).

Step 2) **Characteristics of desired response $G_{yr}(s)$**

The desired input-output characteristics from the reference input $G_{yr}(s)$ is defined by

$$\begin{aligned} G_{yr}(s) &= N(s)H(s) && (12) \\ N(s) &= \frac{I_3}{(1+Ts)^2} \in \mathcal{RH}_\infty && (13) \\ H(s) &= I_3 \in \mathcal{RH}_\infty && (14) \end{aligned}$$

where a nominal model $G(s)$ has a relative degree of 2, and a free parameter of feedforward compensator $H(s)$ is required only for non-minimum phase system. In addition, the feedforward compensator $D(s)$ is determined by

$$D(s) = \frac{s^2}{(1+Ts)^2} I_3 \in \mathcal{RH}_\infty \tag{15}$$

under the coprime factorization of $G(s)$.

The advantage of ILQ-based LQG design is that it enables us to not only specify $G_{yr}(s)$ asymptotically but also adjust $G_{yn}(s)$ easily by changing the tuning parameters of ILQ and/or the observer poles. On the other hand, $\mathcal{H}_\infty$ design method enables us to achieve robust stabilization and hence suppress spillovers.

5 μ-analysis

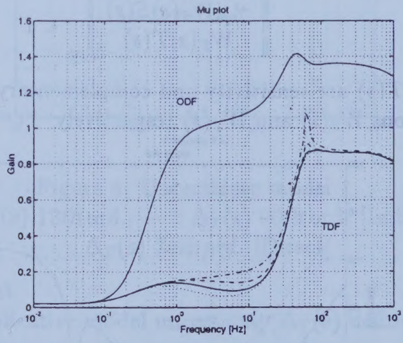


Figure 8: Robust performance with μ-analysis ($T = 0.08$)
— : [ODF], -·-· : $T/4$, - - - : $T/2$, — : T, ··· : $2T$ [TDF]

We finally discuss the results on μ-analysis of the $\mathcal{H}_\infty$ control system with the two degree of freedom (TDF). First of all, it is well known that a robust sensitivity performance of an $\mathcal{H}_\infty$ control system with one degree of freedom (ODF) is satisfied by the main loop theorem [7] as follows:

$$\mu_\Delta(M) < 1 \tag{16}$$

where

$$\begin{aligned} M(s) &= \begin{bmatrix} W_S(s)S(s) & -W_S(s)S(s) \\ W_T(s)T(s) & -W_T(s)T(s) \end{bmatrix} \in \mathcal{RH}_\infty && (17) \\ \Delta(s) &= \mathrm{diag}\{\Delta_S(s), \Delta_T(s)\} \in \mathcal{RH}_\infty && (18) \end{aligned}$$

with $\Delta_S(s) \in \mathcal{BH}_\infty$, $\Delta_T(s) \in \mathcal{BH}_\infty$ and $W_S(s)$ being redefined by $\gamma_{opt}W_S(s)$. In this ODF case, its μ-plots are shown in Fig. 8 where $D(s)$ is $0_{3\times3}$ and $N(s)$ is I_3. From this we see that the robust performance is not achieved because the maximum of μ is 1.41 at the frequencies of the spillovers, although the close loop system is naturally robustly stable. As a result, the levitated vehicle exceeds the maximal measurable gap of the gap-sensors if we supply the reference input of y_1 more than just 0.3 mm. It seems to be quite all right since the important thing in ODF is to suppress the spillovers. That is why we must examine the performance of TDF which is constructed by suitable feedforward dynamics $N(s)$ and $D(s)$ as in Fig. 6. The transfer matrix $M(s)$ is expressed by

$$M(s) = \begin{bmatrix} 0_{3\times3} & -W_S(s)S(s) \\ W_T(s)N(s) & -W_T(s)T(s) \end{bmatrix} \in \mathcal{RH}_\infty \tag{19}$$

and its μ-plots are shown in Fig. 8 for various time constants T of the desired transfer function $G_{yr}(s)$. We find that its robust performance is achieved by making μ less than 1 in frequencies of the spillovers, and the rapid time constant can be selected at the same time since the frequency band width of the desired transfer function $G_{yr}(s)$ is far narrower than that of the complementary function $T(s)$ in the TDF. We have concluded from the viewpoint of practical application that this TDF control system should be constructed in order to satisfy not only robust stability against noise, uncertainties and disturbance but also robust performance with respect to its desired transfer function $G_{yr}(s)$. Note that there is a method known as DK-iteration for achieving the robust performance which is available even in ODF case. In this case, the transient response might be oscillated unless we assign it like TDF. The attitude control, therefore, probably would not be desirable.

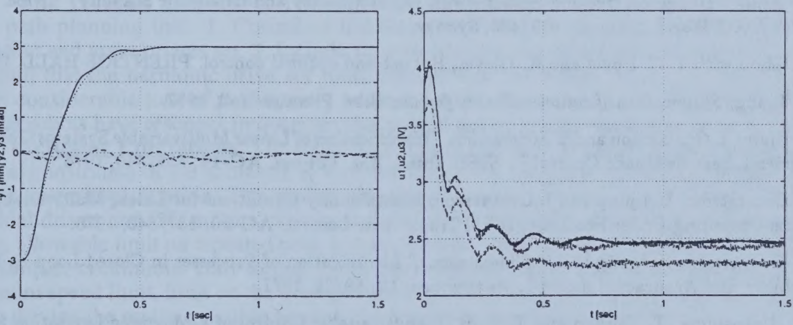


Figure 9: Step response with $\mathcal{H}_\infty$
— : $-y_1$ [mm], u_1 [V], - - - : y_2 [mrad], u_2 [V]
-·-· : y_3 [mrad], u_3 [V], ··· : desired response ($T = 0.08$)

We consider only experimental results of step responses since the spillovers due to an impulse can be suppressed by the $\mathcal{H}_\infty$ controller. The controller have a good and smooth step response as shown in Fig. 9, and it is confirmed that this controller can levitate the vehicle up to the limit of the maximum measurable range of gap-sensors where the magnitude of reference inputs is about 7.3 [mm]. We find that $\mathcal{H}_\infty$ controller might be a little inferior in a strict sense to LQG controller with respect to its tracking ability against reference inputs. In addition, its good decoupling property is also verified by this experiment.

6 Conclusion

In this paper we have discussed a consistent procedure covering from modeling to analysis for suppressing the spillovers with $\mathcal{H}_\infty$ control. First, the exact linearization compensates the nonlinearity of our system in addition to diagonalize, and the closed loop identification estimates the model uncertainties in the modeling. Second, the mixed sensitivity problem designs a robust stabilizing controller in order to suppress the spillovers. Finally, μ-analysis clarifies its robust performance of the two degree of freedom control system. As a future study we try to carry out a performance-preserving frequency weighted controller reduction [16] for the practical procedure since an $\mathcal{H}_\infty$ controller tends to be higher order than a LQG one.

References

[1] IEE-Japan, *Magnetic levitation and magnetic bearings*, Korona Publishing Company, 1993 (in Japanese).

[2] M. Morishita, T. Azukizawa, S. Kanda, N. Tamura and T. Yokoyama, " A New Maglev System for Magnetically Levitated Carrier System " , *IEEE Trans. Vehic. Tech.*, Vol.38, No.4, 230/236, 1989.

[3] T. Fujii, T. Tsujino, K. Suematsu, K. Sasaki and Y. Murata, " Multivariable control of a magnetic levitation system with a Y shape iron plate " , *The 3rd IFAC Symposium on Advances in Control Education*, 1994.

[4] A. Ishidori, *Nonlinear Control Systems — Third Edition*, Springer, 1995.

[5] T. Fujii, " A New Approach to the LQ Design from the Viewpoint of the Inverse Regulator Problem", *IEEE Trans. Aut. Control*, **AC-32**, 995/1004, 1988.

[6] T. Fujii, " Design of Tracking Systems with LQ Optimality and Quadratic Stability " , *Proc. of the 12th IFAC World Congress*, 479/486, Sydney, 1993.

[7] K. Zhou with J. C. Doyle and K. Glover, Robust and optimal control, PRENTICE HALL, 1995.

[8] L. Ljung, *System Identification—Theory for the User*, Prentice-Hall, 1987.

[9] L. Ljung, I. Gustavsson and T.Söderström, "Identification of Linear Multivariable Systems Operating Under Linear Feedback Control " , *IEEE Trans. Aut. Control*, **AC-19**, 836/840, 1974.

[10] T.Söderström, L. Ljung and I. Gustavsson, Identifiability Conditions for Linear Multivariable Systems Operating Under Feedback, IEEE Trans. Aut. Control, **AC-20**, 837/840, 1976.

[11] I. Gustavsson, L. Ljung and T.Söderström, " Identification of Processes in Closed Loop — Identifiability and Accuracy Aspects " , *Automatica*, **13**, 59/75, 1977.

[12] K. Nakashima, T. Tsujino and T. Fujii, " Multivariable Control of a Magnetic Levitation System Using Closed Loop Identification and $\mathcal{H}_\infty$ Control Theory " , *Proc. of the 35th Conf. on Decision and Control*, 3668/3673, Kobe, 1996.

[13] J. Sefton and K. Glover, " Pole/Zero cancellations in the general $\mathcal{H}_\infty$ problem with reference to a two block design " , *Systems & Control letters 14*, 295/306, 1990.

[14] H. Kimura, T. Oike, A. Miura, K. Akai and T. Kida, " Robust stability degree assignment and its applications to the control of the flexible structures " , *Int. J. of Robust and Nonlinear Control*, **1**, 153/169, 1991.

[15] D. J. N. Limebeer, E. M. Kasenally and J. D. Perkins, " On the Design of Robust Two Degree of Freedom Controllers " , *Automatica*, Vol.29, No.1, 157/168, 1993.

[16] P. J. Goddard and K. Glover, " Performance-Preserving Frequency Weighted Controller Approximation: A Coprime Factorization Approach " , *Proc. of the 33rd Conf. on Decision and Control*, 2720/2908, Lake Buena Vista, 1994.

IMPLEMENTING TIME OPTIMAL ROBOT MANEUVERS USING REALISTIC ACTUATOR CONSTRAINTS AND LEARNING CONTROL

Ju Li and Richard W. Longman
Columbia University, New York, New York 10027

Volker H. Schulz
University of Stuttgart, Stuttgart, Germany.

Hans Georg Bock
University of Heidelberg, Heidelberg, Germany

Introduction. This paper is one in a series investigating time optimal control of robots ([1,2] and references [1-7] in [2]). These show that time optimal control can very substantially decrease the time needed to accomplish point-to-point transfers. When applied to robots on an assembly line, increasing the speed of the slowest robot to accomplish its task allows the assembly line to run faster, increasing productivity. The purpose of the current research is to study the issues that must be addressed in order to connect time optimal control to routine path planning in commercial robots using their built in feedback controllers. The research effort is building to a demonstration of increased productivity by use of time optimal robot path planning on a specific Kuka robot in a press chain at the Mercedes Benz manufacturing plant. The development of highly efficient numerical codes for direct solution of optimal control problems, including state inequality constraints, has made it reasonable to aim for routine use of time optimal robot path planning. The code used here is discussed in reference [9] appearing in [2], and other codes that we use appear in [9] of [2], and [5] of [1]. Reference [1] gives an overview of the issues involved in practical implementation on commercial robots. Reference [2] studies the manner in which vibrations should be handled. This paper addresses what the appropriate realistic inequality constraints should be in the optimization, addresses how to handle wear issues, addresses how to implement the control action in hardware using learning control, and combines these results with those of [1,2].

The Overall Objective. The objective is to produce approaches to time optimal robot path planning that: 1. Considers the true set of actuator limitations formulated as a consistent and appropriates set. 2. Include wear or lifetime considerations. People at Mercedes complain that the harmonic drive for joint number 4 of the robots wears out very fast, and causes considerable loss of productivity when the press chain is down for repairs. 3. And the fact that robots have resonant frequencies that should not be unduly excited must be addressed.

Actuator Limitations. The robot time optimal control literature normally models the physical constraints of the actuators as a simple bound on the maximum torque output. This is a gross oversimplification. Examining the manufacturer's manuals for harmonic drives, cycloidal drives and DC motors, we see constraints such as allowable momentary maximum torque, allowable limit on repeated peak torque, limit on allowable acceleration and deceleration peak torque, continuous stall torque limit, rated output torque limit, average torque limit, maximum speed limit, limit on the average speed, possible voltage limit on the power amplifier output driving the motor, possible heating constraints on the motor operation.

A Set of Realistic Actuator/State Constraints. The full paper develops an understanding of what each of these constraints means, and then formulates a consistent set of constraints to impose on the optimization problem. An appropriate maximum torque limit is imposed that is lower than the allowable momentary peak torque limit based on desired lifetime of the drive systems. The rated torque limit is reformulated as an integral constraint on the product of the angular velocity with the torque raised to the 3rd or 10/3 power (depending on the types of bearing used), integrated over the maneuver time. The physical interpretation of this is a limit on the allowed wear per cycle. One can specify a desired drive lifetime such as 10^8 cycles, and from this compute the wear limit. In addition to these limits we impose the instantaneous limit on the angular velocity of each motor, and an integral constraint for the average velocity. The heating constraint is a function of the way in which the motor is mounted, and when needed it forms another integral limit, on the torque squared. The average torque limit suggested by one manufacture of harmonic drives is established as a limit that only applies when operating at maximum speed, and is unreasonably conservative as well as being redundant with respect to the wear limit. Similarly, the motor voltage limit sometimes suggested is an extremely conservative limit ignoring motor back emf, and should not be applied when one wants to optimize performance.

Handling Vibration Limitations. The approach adopted is to perform the optimization on a robot model that does not include flexibility in the drives, since if they are included the open loop optimal trajectory generated may be too sensitive to errors in the modeling. Then the optimization must find an optimal torque input to the robot that is constrained in some way not to excessively excite vibrations. This is consistent with the design of the feedback controllers for commercial robots, that limit the bandwidth of the control to avoid exciting vibrations. The methods used here are those developed in [2].

What to Optimize. We must use the existing commercial robot feedback controllers. One might choose to optimize the command given the feedback controllers, since this is the true control variable we have. Then torque limits are state variable constraints. (The methods of handling the vibrations by picking torques from among smooth functions also do this.) One might also optimize voltage into the motor actuators, or torque seen by the robot. In these two cases, once we obtain the optimized voltage or torque, we must develop a way to determine an associated command to give to the feedback controller. We argue that there is no point to optimizing voltage because of the speed of the motor pole. Back computation methods are developed in [1] to find the command to give a feedback controller based on an optimal torque history and nominal feedback controller model. Or we can use learning control.

Use of Learning Control to Determine Command. We introduce the concept of using learning control to perform the task of making the hardware execute the trajectory that we have determined as the optimal trajectory. This addresses an important issue. One cannot simply give the desired trajectories of the links as commands to the feedback controllers, because they will not actually perform the command requested. One will loose much of the time advantage of time optimal control because the feedback control system is always behind in its execution of the command. Also, it is important to have some way of fixing errors in the computed torque histories produced by inaccuracies in the model used. Such errors will result in having the robot fail to exactly reach the final point on the trajectory, and perhaps fail to reach zero velocity at the desired time and place.

During the learning process, it is quite possible that some of the inequality constraints will be violated. These inequalities are not hard inequalities, and it is OK to violate them 10 or 20 times, if the violation is not too extreme. For example, violating the wear limit means that one develops more wear per cycle during this violation. If done repeatedly, this would reduce the lifetime from say 10^8 cycles down to 10^5 cycles, but done 20 times produces negligible effect. During the learning process, one should monitor any violation. It is also possible that once convergence is reached one still violates some. If after learning is complete, one determines a constraints is violated, then one can return to the optimization problem, and lower the value of the constraint imposed, to leave some extra margin to model inaccuracies, and then try again. One can do the same during the learning, and one can also perform a homotopy at the end to reach the desired limit. We believe this approach is the most practical for handling both the issues of inaccuracies in the model of the robot, and for eliminating the need for back calculation of the command from optimal torque and joint variable histories, and subsequent adjustment for the results of any model errors. In addition, we know we have model error, because we have chosen to optimize based on a robot model without flexibility in the joints, and we apply it to the real robot which does have flexibility. This learning approach is a reasonable engineering solution to this discrepancy. Note that one needs to extend the desired trajectory at the end of the maneuver, making the flexible model learn to sit at the final endpoint without residual vibrations. Illustrations of the optimization process, the choices for handling vibrations, and the application of learning control are given in the full paper.

References.

1. J. Li, R. W. Longman, M. Steinbach, and H. G. Bock, "Issues in the Implementation of Time-Optimal Path Planning," *Proceedings of the Symposium on Nonlinear Dynamics*, AIAA Aerospace Sciences Meeting, Reno, Nevada, January 1997.

2. J. Li, R. W. Longman, V. H. Schulz, and H. G. Bock, "The Choice of Appropriate Cost Functionals for Optimizing the Operating Speed of Robots," AAS/AIAA Astrodynamics Conference, Sun Valley, Idaho, August 1997, to appear in the *Advances in the Astronautical Sciences.*

redundant with respect to the wear limit. Similarly, the motor voltage limit sometimes suggested is an extremely conservative limit ignoring motor back emf, and should not be applied when one wants to optimize performance.

Handling Vibration Limitations. The approach adopted is to perform the optimization on a robot model that does not include flexibility in the drives, since if they are included the open loop optimal trajectory generated may be too sensitive to errors in the modeling. Then the optimization must find an optimal torque input to the robot that is constrained in some way not to excessively excite vibrations. This is consistent with the design of the feedback controllers for commercial robots, that limit the bandwidth of the control to avoid exciting vibrations. The methods used here are those developed in [2].

What to Optimize. We must use the existing commercial robot feedback controllers. One might choose to optimize the command given the feedback controllers, since this is the true control variable we have. Then torque limits are state variable constraints. (The methods of handling the vibrations by picking torques from among smooth functions also do this.) One might also optimize voltage into the motor actuators, or torque seen by the robot. In these two cases, once we obtain the optimized voltage or torque, we must develop a way to determine an associated command to give to the feedback controller. We argue that there is no point to optimizing voltage because of the speed of the motor pole. Back computation methods are developed in [1] to find the command to give a feedback controller based on an optimal torque history and nominal feedback controller model. Or we can use learning control.

Use of Learning Control to Determine Command. We introduce the concept of using learning control to perform the task of making the hardware execute the trajectory that we have determined as the optimal trajectory. This addresses an important issue. One cannot simply give the desired trajectories of the links as commands to the feedback controllers, because they will not actually perform the command requested. One will loose much of the time advantage of time optimal control because the feedback control system is always behind in its execution of the command. Also, it is important to have some way of fixing errors in the computed torque histories produced by inaccuracies in the model used. Such errors will result in having the robot fail to exactly reach the final point on the trajectory, and perhaps fail to reach zero velocity at the desired time and place.

During the learning process, it is quite possible that some of the inequality constraints will be violated. These inequalities are not hard inequalities, and it is OK to violate them 10 or 20 times, if the violation is not too extreme. For example, violating the wear limit means that one develops more wear per cycle during this violation. If done repeatedly, this would reduce the lifetime from say 10^8 cycles down to 10^5 cycles, but done 20 times produces negligible effect. During the learning process, one should monitor any violation. It is also possible that once convergence is reached one still violates some. If after learning is complete, one determines a constraints is violated, then one can return to the optimization problem, and lower the value of the constraint imposed, to leave some extra margin to model inaccuracies, and then try again. One can do the same during the learning, and one can also perform a homotopy at the end to reach the desired limit. We believe this approach is the most practical for handling both the issues of inaccuracies in the model of the robot, and for eliminating the need for back calculation of the command from optimal torque and joint variable histories, and subsequent adjustment for the results of any model errors. In addition, we know we have model error, because we have chosen to optimize based on a robot model without flexibility in the joints, and we apply it to the real robot which does have flexibility. This learning approach is a reasonable engineering solution to this discrepancy. Note that one needs to extend the desired trajectory at the end of the maneuver, making the flexible model learn to sit at the final endpoint without residual vibrations. Illustrations of the optimization process, the choices for handling vibrations, and the application of learning control are given in the full paper.

References.

1. J. Li, R. W. Longman, M. Steinbach, and H. G. Bock, "Issues in the Implementation of Time-Optimal Path Planning," *Proceedings of the Symposium on Nonlinear Dynamics*, AIAA Aerospace Sciences Meeting, Reno, Nevada, January 1997.
2. J. Li, R. W. Longman, V. H. Schulz, and H. G. Bock, "The Choice of Appropriate Cost Functionals for Optimizing the Operating Speed of Robots," AAS/AIAA Astrodynamics Conference, Sun Valley, Idaho, August 1997, to appear in the *Advances in the Astronautical Sciences*.

Fundamental Limitations of Disturbance Rejection in Feedback Control Systems

Daniel E. Davison Pierre T. Kabamba Semyon M. Meerkov[†]

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI, 48109-2122

Abstract—This paper addresses the issue of limitations of disturbance rejection that are a consequence of the Bode sensitivity integral and of design specifications that can be written as constraints on the sensitivity function. The disturbance is modeled as a stationary random process with known power spectral density, and disturbance rejection is measured by the output variance. The design specifications considered include constraints on the closed-loop system bandwidth, constraints on the stability robustness, and constraints on the low-frequency tracking. It is shown that these constraints place lower bounds on the achievable output variance. Analytical expressions for the lower bounds are obtained, and the bounds are applied to an autopilot design problem for a linearized model of an aircraft in turbulence.

1 Introduction

In any feedback control problem, there are fundamental constraints that dictate what the achievable levels of performance are. Probably the most well known constraint is the algebraic identity $S + T = 1$, where $S \triangleq 1/(1 + PC)$ and $T \triangleq PC/(1 + PC)$ are, respectively, the sensitivity and complementary sensitivity functions for the classical feedback loop in Figure 1. This constraint, which is due to the topology of the feedback loop, implies that it is impossible to reject a particular frequency component of both the disturbance, d, and the sensor noise, n, simultaneously. Thus, if d and n have power at common frequencies, it is *impossible* to design a linear time-invariant finite-dimensional (LTIFD) controller that has a noise-free output signal—this is a fundamental limitation. Other important mathematical constraints exist. For example, if PC has relative degree greater than one, then

$$\int_0^\infty \ln |S(j\omega)|\, d\omega = c \triangleq \pi \sum_{i=1}^{N_p} \operatorname{Re} p_i \geq 0, \tag{1}$$

where $\{p_1, \ldots, p_{N_p}\}$ are the unstable poles of PC. This is the so-called Bode sensitivity integral, derived in the above form in [1]. It is well documented (e.g., see [1]–[6]) that (1) represents a conservation law on sensitivity (or, "dirt", as coined by Gunter Stein

---

[†] Please send all correspondence to Professor Semyon M. Meerkov, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48109-2122, U.S.A., phone +1 (313) 763-6349, fax +1 (313) 763-8041, email `smm@eecs.umich.edu`.

This work was supported by the Army Research Office under Grant DAAL03-92-G-0127 and by graduate student fellowships awarded by the Natural Sciences and Engineering Research Council of Canada and the Rackham School of Graduate Studies at the University of Michigan.

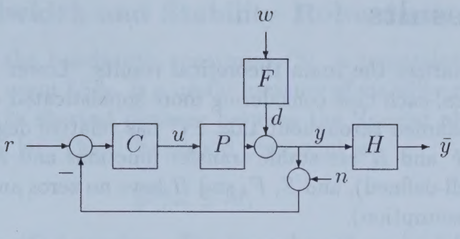


Figure 1: System Model.

[2]): Any reduction in sensitivity in one frequency range is necessarily countered by a sensitivity increase in another frequency range. Moreover, if the closed-loop bandwidth is constrained (as it always is in practice), then the "dirt" cannot be placed outside the bandwidth, and, therefore, demanding sufficiently large sensitivity reduction in any given frequency range will lead to an unacceptable peak in the sensitivity function in some other frequency range, i.e., the norm $\|S\|_\infty$ will be large. Such limitations are discussed in detail in [1]. Although $\|S\|_\infty$ is an important measure of system performance (e.g., it is a measure of stability robustness and it is the worst-case L_2-norm of y when d is constrained to have unity L_2-norm), it often is not the most appropriate measure of disturbance rejection. For example, if d is a stationary random process with known power spectral density (modeling turbulence acting on an airplane or water waves acting on a ship, for example), then the variance of y is often a more useful measure since it provides information on the expected deviation in y from its nominal value. This paper addresses the issue of limitations on the steady-state variance of $\bar{y}$ (the filtered version of y shown in Figure 1) that are a consequence of (1), when subjected to bandwidth (and other) design specifications. To the best of the authors' knowledge, this particular problem has not been investigated previously.

As indicated in Figure 1, it is assumed that the disturbance is generated by filtering w, standard white noise, through stable filter F. It is also assumed that sensor noise is negligible within the closed-loop system bandwidth, where the notion of bandwidth will be precisely defined later. Lastly, to ensure that the variance of $\bar{y}$ is finite and that (1) applies, it is assumed that FH is stable and strictly proper and that PC has relative degree greater than one. Under these assumptions, the steady-state variance of $\bar{y}$ is $\|SFH\|_2^2$. Therefore, the problem is to investigate what (1) implies about the achievable value of $\|SFH\|_2^2$. The main results of this investigation are given in Section 2. In particular, lower bounds on $\|SFH\|_2^2$ are given under progressively complicated sets of constraints, starting with only a constraint on bandwidth, and ending with a constraint on bandwidth and a general constraint on $|S(j\omega)|$. An example of the use of the resulting bounds is given in Section 3, where an aircraft subjected to turbulence is considered. Conclusions are drawn in Section 4 and proofs are given in the Appendix.

Remark 1.1 Constant c in (1) depends on the unstable poles of both the controller and the plant. Since the controller poles are not known before the controller is chosen, one might conclude that the bounds derived in this paper are of no use. Fortunately, this is not the case because the bounds are all increasing functions of c, and, therefore, if c is chosen under the assumption that the controller has no unstable poles, then the resulting lower bounds remain valid lower bounds even if the controller does, in fact, have unstable poles. (It also follows that the lower bounds given here are not always achievable.) □

2 Main Results

This section summarizes the main theoretical results. Lower bounds on $\|SFH\|_2^2$ are derived in four cases, each case considering more sophisticated constraints than the previous case. It is assumed throughout that PC has relative degree greater than one (so that (1) applies), F and H are stable transfer functions and FH is strictly proper (so that $\|SFH\|_2$ is well-defined), and S, F, and H have no zeros and poles on the imaginary axis (a technical assumption).

2.1 Case 1: Bandwidth Constraint

The notion of *bandwidth* that we desire to capture mathematically is that frequency, Ω, above which, for all practical purposes, no control effort is desired. The value of Ω depends on sensor noise frequency distribution, actuator limitations, and the desire to avoid exciting high frequency plant dynamics, among other things. The simplest method of expressing this notion of bandwidth is to restrict the class of controllers to those LTIFD controllers satisfying $\forall\omega > \Omega,\ C(j\omega) = 0$, or, equivalently,

$$\forall\omega > \Omega, \quad S(j\omega) = 1. \tag{2}$$

Strictly speaking, this constraint never holds, except for the case where C is identically zero; it is possible to loosen the constraint by allowing $|S(j\omega) - 1|$ to be small for $\omega > \Omega$ (and tending to zero as $\omega \to \infty$), but this complicates the notation without adding much insight and without changing the results significantly. Under bandwidth constraint (2), the following bound holds:

Theorem 2.1 Let C be any LTIFD internally-stabilizing controller whose corresponding sensitivity function satisfies (2). Define $c \geq 0$ as in (1) and define the constant

$$\alpha = \frac{2c}{\Omega} + \frac{1}{\Omega}\int_0^{\Omega} \ln|F(j\omega)H(j\omega)|^2\,d\omega.$$

Then the following inequality holds:

$$\|SFH\|_2^2 \geq \frac{\Omega}{\pi}e^{\alpha} + \frac{1}{\pi}\int_{\Omega}^{\infty} |F(j\omega)H(j\omega)|^2\,d\omega. \tag{3}$$

The bound in (3) is achieved by the (possibly unrealizable) sensitivity function that satisfies

$$\forall\omega \in [0,\Omega], \quad |S^*(j\omega)| = \frac{1}{|F(j\omega)H(j\omega)|}e^{\alpha/2}. \tag{4}$$

Proof: See the Appendix. □

Some remarks are in order. First note that the bound in (3) can always be evaluated in closed-form, but it is easier to evaluate it numerically. Second, the function $|S^*|$ in the above theorem is the shape of the sensitivity function that results in equality in (3). However, there does not exist a controller such that $|S|$ is exactly equal to $|S^*|$ for all ω because, exact equality in (2) is generally not possible, $|S^*|$ is discontinuous at $\omega = \Omega$, and non-minimum-phase zeros put interpolation constraints on S that have not been accounted for. Nevertheless, $|S^*|$ is included since it gives insight into what the shape of $|S|$ should be to result in minimum $\|SFH\|_2^2$. Finally, note that the bound is a continuous function of Ω. As $\Omega \to 0$, the bound tends to $\|FH\|_2^2$ if $c = 0$ and to ∞ if $c > 0$. These results are expected since, at $\Omega = 0$, the only admissible controller is $C = 0$, and the corresponding value of $\|SFH\|_2^2$ is $\|FH\|_2^2$ if P is stable and ∞ if P is unstable.

2.2 Case 2: Bandwidth and Stability Robustness Constraints

Suppose, in addition to the bandwidth constraint (2), a constraint is placed on the stability robustness. The norm $\|S\|_\infty$ is a useful measure of stability robustness and has a simple interpretation: the shortest distance between the Nyquist plot and the critical point, $(-1, j0)$, is $1/\|S\|_\infty$ [4]. Thus, the second constraint is

$$\|S\|_\infty \leq M, \tag{5}$$

where $M > 1$ is a prespecified number. The lower bound on $\|SFH\|_2^2$ is modified as follows:

Theorem 2.2 Let C be any LTIFD internally-stabilizing controller whose corresponding sensitivity function satisfies both (2) and (5). Define $c \geq 0$ as in (1), and assume that $\Omega \ln M \geq c$. Define $\beta \geq 0$ to be the unique scalar in the range

$$\left[\min_{\omega \in [0,\Omega]} |F(j\omega)H(j\omega)|,\ M \cdot \max_{\omega \in [0,\Omega]} |F(j\omega)H(j\omega)|\right]$$

satisfying

$$\int_0^\Omega \ln\left\{\min\left(M, \frac{\beta}{|F(j\omega)H(j\omega)|}\right)\right\} d\omega = c. \tag{6}$$

Then

$$\|SFH\|_2^2 \geq \frac{1}{\pi}\int_0^\Omega \min(M^2|F(j\omega)H(j\omega)|^2, \beta^2)\, d\omega + \frac{1}{\pi}\int_\Omega^\infty |F(j\omega)H(j\omega)|^2\, d\omega. \tag{7}$$

The bound in (7) is achieved by the (possibly unrealizable) sensitivity function that satisfies

$$\forall \omega \in [0, \Omega], \quad |S^*(j\omega)| = \min\left(M, \frac{\beta}{|F(j\omega)H(j\omega)|}\right). \tag{8}$$

Proof: See the Appendix. ☐

Note that the condition $\Omega \ln M \geq c$ stated in the theorem is necessary for the existence of a sensitivity function S that satisfies all of (1), (2), and (5). In addition, bound (7) is a continuous decreasing function of M and, as $M \to \infty$, the bound of Theorem 2.1 is recovered. In terms of computation, because the left-hand-side of (6) is a monotonic function of β, the bisection algorithm can be used to solve for β.

2.3 Case 3: Bandwidth, Stability Robustness, and Command-Tracking Constraints

Besides the bandwidth and stability robustness constraints, there are often specifications on the command tracking behavior of the closed-loop system. Since the transfer function from reference to output error is S, such specifications can often be stated as constraints on $|S|$. Suppose that the command-tracking constraint is of the form

$$\forall \omega \in [0, \omega_1], \quad |S(j\omega)| \leq m, \tag{9}$$

where $\omega_1 \geq 0$ and $m \in (0, M)$ are prespecified real numbers. If this constraint is applied, in addition to the bandwidth and stability robustness constraints of Case 2, then the following lower bound on $\|SFH\|_2^2$ exists:

Theorem 2.3 Let C be any LTIFD internally-stabilizing controller whose corresponding sensitivity function satisfies (2), (5), and (9). Define $c \geq 0$ as in (1), and assume that $\omega_1 \ln m + (\Omega - \omega_1) \ln M \geq c$. Define $\beta \geq 0$ to be the unique scalar in the range

$$\left[\min_{\omega \in [0,\Omega]} |F(j\omega)H(j\omega)|,\ \max\left\{ m \cdot \max_{\omega \in [0,\omega_1]} |F(j\omega)H(j\omega)|,\ M \cdot \max_{\omega \in [\omega_1,\Omega]} |F(j\omega)H(j\omega)| \right\} \right]$$

satisfying

$$\int_0^{\omega_1} \ln \left\{ \min \left(m, \frac{\beta}{|F(j\omega)H(j\omega)|} \right) \right\} d\omega + \int_{\omega_1}^{\Omega} \ln \left\{ \min \left(M, \frac{\beta}{|F(j\omega)H(j\omega)|} \right) \right\} d\omega = c. \tag{10}$$

Then

$$\begin{aligned} \|SFH\|_2^2 \ \geq\ & \frac{1}{\pi} \int_0^{\omega_1} \min(m^2 |F(j\omega)H(j\omega)|^2, \beta^2)\, d\omega \\ & + \frac{1}{\pi} \int_{\omega_1}^{\Omega} \min(M^2 |F(j\omega)H(j\omega)|^2, \beta^2)\, d\omega + \frac{1}{\pi} \int_{\Omega}^{\infty} |F(j\omega)H(j\omega)|^2\, d\omega. \end{aligned} \tag{11}$$

The bound in (11) is achieved by the (possibly unrealizable) sensitivity function that satisfies

$$\forall \omega \in [0, \Omega], \quad |S^*(j\omega)| = \begin{cases} \min\left(m, \beta/|F(j\omega)H(j\omega)|\right) & \text{if} \quad 0 \leq \omega \leq \omega_1 \\ \min\left(M, \beta/|F(j\omega)H(j\omega)|\right) & \text{if} \quad \omega_1 < \omega \leq \Omega. \end{cases} \tag{12}$$

Proof: See the Appendix. □

The condition $\omega_1 \ln m + (\Omega - \omega_1) \ln M \geq c$ in the theorem is necessary for the existence of a sensitivity function S that satisfies all of (1), (2), (5), and (9). Also note that, when $m = M$, the Theorem 2.2 bound is recovered; as m and M tend to ∞, the Theorem 2.1 bound is recovered. As before, the bisection algorithm can be used to solve for β.

2.4 Case 4: Bandwidth and General $|S|$ Constraints

Suppose that, in addition to the bandwidth constraint (2), the following general constraint on $|S|$ is specified:

$$\forall \omega \in [0, \Omega], \ |S(j\omega)| \leq M_{\max}(\omega), \tag{13}$$

where $M_{\max}(\cdot)$ is a prespecified positive function of $\omega \in [0, \Omega]$. The constraints on $|S|$ in Cases 1, 2, and 3 are all particular instances of (13). The following lower bound on $\|SFH\|_2^2$ holds:

Theorem 2.4 Let C be any LTIFD internally-stabilizing controller whose corresponding sensitivity function satisfies (2) and (13). Define $c \geq 0$ as in (1), and assume that

$$\int_0^{\Omega} \ln M_{\max}(\omega)\, d\omega \geq c. \tag{14}$$

Define $\beta \geq 0$ to be the unique scalar in the range

$$\left[\min_{\omega \in [0,\Omega]} |F(j\omega)H(j\omega)|,\ \max_{\omega \in [0,\Omega]} M_{\max}(\omega) |F(j\omega)H(j\omega)| \right]$$

satisfying

$$\int_0^\Omega \ln\left\{\min\left(M_{\max}(\omega), \frac{\beta}{|F(j\omega)H(j\omega)|}\right)\right\} d\omega = c. \tag{15}$$

Then

$$\|SFH\|_2^2 \geq \frac{1}{\pi}\int_0^\Omega \min(M_{\max}^2(\omega)|F(j\omega)H(j\omega)|^2, \beta^2)\, d\omega + \frac{1}{\pi}\int_\Omega^\infty |F(j\omega)H(j\omega)|^2\, d\omega. \tag{16}$$

The bound in (16) is achieved by the (possibly unrealizable) sensitivity function that satisfies

$$\forall\omega \in [0,\Omega], \quad |S^*(j\omega)| = \min\left(M_{\max}(\omega), \frac{\beta}{|F(j\omega)H(j\omega)|}\right). \tag{17}$$

Proof: See the Appendix. □

3 Application to an Aircraft in Turbulence

In this section we consider an F-89 jet, cruising at 20000 ft at Mach 0.638, under autopilot control. The autopilot's main purpose is to track the reference altitude (20000 ft) despite the presence of turbulence; however, the autopilot is constrained to meet specifications on bandwidth, stability robustness, and passenger comfort. For simplicity, assume that the aircraft is restricted to fly in a vertical plane and that the autopilot uses only the elevator for control.

In the following subsections, the airplane and turbulence models are given and the design specifications are stated precisely. Then, the tools from Section 2 are used to investigate the relationship between the attainable level of passenger comfort and the closed-loop system bandwidth.

Models

Using the information in [7], the transfer functions from elevator angle deviation (u, in rad) to altitude deviation (y, in ft) and vertical turbulence velocity (w_g, in ft/s) to altitude deviation can be computed to be

$$\begin{aligned}\frac{y(s)}{u(s)} &= \frac{69.8(s+19.8114)(s-17.0301)(s+0.0064)}{s(s^2+4.2087s+18.2546)(s^2+0.0090s+0.0040)},\\ \frac{y(s)}{w_g(s)} &= \frac{-1.43(s+3.8443)(s^2+0.0055s+0.0131)}{s(s^2+4.2087s+18.2546)(s^2+0.0090s+0.0040)}.\end{aligned}$$

The turbulence is modeled as in [8]. Specifically, w_g is modeled as a stationary random process with power spectral density

$$\sigma_w^2 \frac{L}{V}\frac{4}{4+(L\omega/V)^2}, \tag{18}$$

where V is the airspeed of the aircraft (660 ft/s), L is the turbulence scale, and σ_w is the turbulence intensity. According to [9], suitable values for L and σ_w (at 20000 ft altitude, and assuming "clear air" turbulence) are, respectively, 1750 ft and 4.8 ft/s. A random process with spectrum (18) can be generated by filtering white noise with the filter

$$\sigma_w\sqrt{\frac{L}{V}}\frac{2}{2+\frac{L}{V}s} = \frac{5.90}{s+0.754}.$$

Qualitative comfort description	Variance of vertical acceleration (in $(\mathrm{ft/s^2})^2$)
Negligible	2.59
Slight	10.37
Moderate	10.37–23.33
Moderately heavy	41.47–93.32
Severe	93.32–373.3
Extreme	373.3

Table 1: Empirical relationship between airplane travel comfort and variance of vertical acceleration.

These plant and turbulence models fit into Figure 1 with

$$\begin{aligned} P(s) &= \frac{69.8(s+19.8114)(s-17.0301)(s+0.0064)}{s(s^2+4.2087s+18.2546)(s^2+0.0090s+0.0040)}, \\ F(s) &= \frac{-8.44(s+3.8443)(s^2+0.0055s+0.0131)}{s(s+0.754)(s^2+4.2087s+18.2546)(s^2+0.0090s+0.0040)}. \end{aligned}$$

Signal u is the elevator deviation, y is the altitude deviation, n is high-frequency sensor noise, and $\ddot{y}$ is the vertical acceleration, i.e., $H(s) = s^2$. Signal d is the equivalent output disturbance; it is not w_g, the turbulence velocity. Also note that F is unstable, but the pole at $s = 0$ is harmless because S necessarily has a zero at $s = 0$.

Design Specifications

S1: To ensure stability robustness, the Nyquist plot should remain at least 1/2 units away from the critical point, that is, $\|S\|_\infty \leq 2$;

S2: To ensure that the autopilot can track slowly-varying reference signals, the sensitivity function must satisfy $|S(j\omega)| \leq -40\,\mathrm{dB}$ for all $\omega \leq 0.5\,\mathrm{rad/s}$;

S3: Due to sensor and actuator limitations (including sensor noise), the closed-loop bandwidth is required to be no more than Ω rad/s; and

S4: Passenger comfort, as measured by the variance of $\ddot{y}$, must be satisfactory. (Table 1, taken from [10], shows the relationship between comfort and variance of $\ddot{y}$.) Specifically, the variance of $\ddot{y}$ must be less than γ $(\mathrm{ft/s^2})^2$.

Analysis

The tools from Section 2 can be used to analyze tradeoffs and performance limitations. In this subsection we analyze one such situation, namely, the tradeoff between passenger comfort and bandwidth (i.e., γ and Ω). The parameter values used below are $c = 0$ and (from Specifications S1 and S2) $m = 0.01$ (i.e., $-40\,\mathrm{dB}$), $\omega_1 = 0.5\,\mathrm{rad/s}$, and $M = 2$.

First observe that, subject to Specifications S1–S3, (11) is a valid lower bound on the achievable variance of $\ddot{y}$, i.e., a lower bound on the achievable passenger comfort. A plot of (11), as a function of Ω, is given in Figure 2. (Note that the curve is drawn only for $\Omega \geq \omega_1 + (c - \omega_1 \ln m)/\ln M = 3.82\,\mathrm{rad/s}$, a consequence of the requirement that Specifications S1–S3 be consistent with constraint (1), as stated in Theorem 2.3.) The plot provides useful information about the relationship between γ and Ω. For example, suppose it is required that, in clear air turbulence, the worst level of discomfort permissible is "slight." From Table 1, the corresponding value of γ is $10.37\,(\mathrm{ft/s^2})^2$. From Figure 2, the corresponding value of Ω is 18.5 rad/s. In other words, to satisfy Specifications S1, S2,

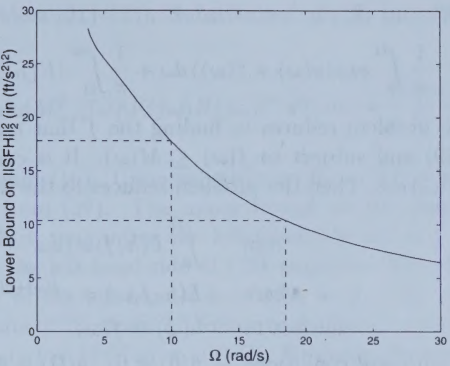


Figure 2: A lower bound on the variance of $\ddot{y}$ subject to Specifications S1–S3.

and S4 (with $\gamma = 10.37$), the closed-loop bandwidth *must be* at least 18.5 rad/s. Knowing such a lower bound on the bandwidth, in turn, determines, among other things, the minimum quality of sensors and actuators. Alternatively, suppose that, instead of fixing Specification S4, Specification S3 is defined (perhaps due to an a priori selection of actuators and sensors) by setting $\Omega = 10$ rad/s. From Figure 2, the corresponding lower bound on γ is 17.8 (ft/s^2)^2, which, from Table 1, corresponds to "moderate" discomfort. In other words, subject to Specifications S1–S3 (with $\Omega = 10$ rad/s), it is *impossible* to attain a passenger comfort level better than "moderate"—this is a fundamental limitation on disturbance rejection.

Note that the assumption that F and P have no poles or zeros on the imaginary axis is violated in this example. However, it can be shown that the results of Theorem 2.3 are valid if the imaginary axis poles and zeros are shifted slightly into the left-half-plane.

4 Conclusions

In this paper it was shown how constraining the closed-loop system bandwidth and placing upper bounds on the sensitivity function limit the ability of a controller to reject disturbances. The results were applied to an aircraft example; in particular, a tradeoff curve between the best achievable output variance and the closed-loop system bandwidth, subject to stability robustness and low-frequency tracking specifications, was presented.

Appendix: Proof of Theorems 2.1, 2.2, 2.3, and 2.4

The proof for Theorem 2.4 is presented first, then Theorems 2.1, 2.2, and 2.3 are proven by using particular forms for function $M_{\max}$. The key idea in the proof of Theorem 2.4 is to treat the function $|S(j\omega)|$ as a curve in $\omega \in [0, \Omega]$, and apply classical optimization theory to find the S curve that minimizes $\|SFH\|_2^2$. The optimization is carried out subject to constraints (1) and (13). To set up the optimization problem, define the functions $f(\omega) = \ln|S(j\omega)|^2$, $g(\omega) = \ln|F(j\omega)H(j\omega)|^2$, and $M(\omega) = \ln M_{\max}^2(\omega)$. Use (1) and the assumption that $|S(j\omega)| = 1$ for $\omega > \Omega$ to obtain

$$\int_0^{\Omega} f(\omega)\, d\omega = 2c. \tag{19}$$

Now compute

$$\|SFH\|_2^2 = \frac{1}{\pi}\int_0^{\Omega} \exp(g(\omega)+f(\omega))\,d\omega + \frac{1}{\pi}\int_{\Omega}^{\infty} |F(j\omega)H(j\omega)|^2\,d\omega. \tag{20}$$

Thus, the optimization problem reduces to finding the f that minimizes $\int_0^{\Omega}\exp(g(\omega)+f(\omega))\,d\omega$ subject to (19) and subject to $f(\omega) \leq M(\omega)$. It is convenient to define, for $\omega \in [0,\Omega]$, $h(\omega) = \int_0^{\omega} f(\bar{\omega})\,d\bar{\omega}$. Then the problem reduces to the following subproblem:

$$\begin{aligned} &\min_f && \int_0^{\Omega} L(h,f,\omega)\,d\omega && (21)\\ &\text{where} && L(h,f,\omega) = e^{f(\omega)+g(\omega)} && (22)\\ &\text{subject to} && \dot{h}(\omega) = f(\omega) && (23)\\ &\text{with end conditions} && h(0)=0, \quad h(\Omega) = 2c && (24)\\ &\text{and control constraint} && f(\omega) \leq M(\omega). && (25)\end{aligned}$$

Here, f plays the role of the control signal and h plays the role of the state. The "dot" notation denotes differentiation with respect to ω. The solution of (21)-(25) can be found by applying Pontryagin's Minimum Principle. The Hamiltonian is $H(h,f,\omega) = L(h,f,\omega) + \lambda f = \exp(f+g) + \lambda f$, where λ is the costate, i.e., the solution to $-\dot{\lambda} = \partial H/\partial h$. Since $\partial H/\partial h = 0$, it follows that λ is constant. Pontryagin's Minimum Principle states that the optimal control, $f(\omega)$, is that which minimizes the Hamiltonian (when evaluated at the optimal state and costate) over all admissible controls. Here, this means that the optimal control is the $f(\omega)$ that minimizes

$$\exp(f(\omega)+g(\omega)) + \lambda f(\omega) \tag{26}$$

where λ is the optimal costate (yet to be determined) and where $f(\omega)$ is constrained by $f(\omega) \leq M(\omega)$. This holds at each $\omega \in [0,\Omega]$. If $\lambda \geq 0$, then function (26) is clearly "minimized" at $f(\omega) = -\infty$. Since λ is a constant, this implies that $f(\omega) = -\infty$ for all $\omega \in [0,\Omega]$. However, such a control signal does not satisfy (19), so it must be that $\lambda < 0$. Simple analysis shows that, for $\lambda < 0$, (26), as a function of $f(\omega)$, has exactly one stationary point (a global minimum) at $f(\omega) = \ln(-\lambda) - g(\omega)$. Therefore, the optimal control, in terms of the still unknown λ, is

$$f(\omega) = \min(M(\omega), \ln(-\lambda) - g(\omega)) = \min(M(\omega), -g(\omega)+c_1), \tag{27}$$

where $c_1 = \ln(-\lambda)$. We can solve for c_1 by applying constraint (19), yielding

$$\int_0^{\Omega} \min(M(\omega), -g(\omega)+c_1)\,d\omega = 2c. \tag{28}$$

Equation (28) can be solved for c_1 via the bisection algorithm since the right-hand side is a known quantity and the left-hand side is a monotonically increasing function of c_1. From the monotonicity of the left-hand-side of (28) in c_1 and from the observations that the maximum of the left-hand-side of (28) (as a function of c_1) is $\int_0^{\Omega} M(\omega)\,d\omega$ and that the minimum of the left-hand-side of (28) is $-\infty$, it follows that a necessary and sufficient condition for the existence of a solution to (28) is that $\int_0^{\Omega} M(\omega)\,d\omega \geq 2c$. The optimal cost is

$$\begin{aligned}\int_0^{\Omega} \exp\{f(\omega)+g(\omega)\}\,d\omega &= \int_0^{\Omega} \exp\{\min(M(\omega), -g(\omega)+c_1) + g(\omega)\}\,d\omega\\ &= \int_0^{\Omega} \exp\{\min(M(\omega)+g(\omega), c_1)\}\,d\omega. \qquad (29)\end{aligned}$$

This solves the subproblem (21)–(25). Substitution of (29) into (20) yields, after simple manipulation,

$$\|SFH\|_2^2 \geq \frac{1}{\pi}\int_0^{\Omega} \min(M_{\max}^2(\omega)|F(j\omega)H(j\omega)|^2, e^{c_1})\,d\omega + \frac{1}{\pi}\int_{\Omega}^{\infty} |F(j\omega)H(j\omega)|^2\,d\omega.$$

Define $\beta \triangleq e^{c_1/2}$ to obtain (16). Upon substitution for g, M, and c_1, (15) follows from (28) and (17) follows from (27). The upper bound on the range where β lies is the smallest value of β which maximizes the left-hand-side of (15); the lower bound is a value of β that makes the left-hand-side of (15) negative. The uniqueness of β follows from the monotonicity of the left-hand-side of (15) in β. This completes the proof of Theorem 2.4. The results of Theorem 2.3 follow immediately upon setting $M_{\max}(\omega) = m$ (if $\leq \omega \leq \omega_1$) and $M_{\max}(\omega) = M$ (if $\omega_1 < \omega \leq \Omega$). Similarly, the results from Theorem 2.2 follow immediately using $M_{\max}(\omega) = M$ (for all $0 < \omega \leq \Omega$). Finally, the results from Theorem 2.1 follow by letting $M_{\max}(\omega)$ tend to ∞ for each $\omega \in [0, \Omega]$. In particular, the relationship between β (in Theorem 2.4) and α (in Theorem 2.1) is $\beta = \exp(\alpha/2)$.

References

[1] J. S. Freudenberg and D. P. Looze, "Right half-plane poles and zeros and design tradeoffs in feedback systems," *IEEE Transactions on Automatic Control*, vol. AC-30, pp. 555–565, June 1985.

[2] G. Stein, "Respect the unstable," in *Proceedings of the 1989 Conference for Decision and Control*, (Tampa, FL), IEEE, Dec. 1989. (Hendrik W. Bode Lecture).

[3] D. P. Looze and J. S. Freudenberg, "Limitations of feedback properties imposed by open-loop right half plane poles," *IEEE Transactions on Automatic Control*, vol. 36, pp. 736–739, June 1991.

[4] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback Control Theory*. New York, NY: Macmillan, 1992.

[5] D. P. Looze and J. S. Freudenberg, "Tradeoffs and limitations in feedback systems," *The Control Handbook*, pp. 537–550, 1996.

[6] M. M. Seron, J. H. Braslavsky, and G. C. Goodwin, *Fundamental Limitations in Filtering and Control*. Comm. and Control Engineering, London: Springer, 1997.

[7] D. McRuer, I. Ashkenas, and D. Graham, *Aircraft Dynamics and Automatic Control*. Princeton, New Jersey: Princeton University Press, 1973.

[8] J. R. Elliott, "NASA's advanced control law program for the F-8 digital fly-by-wire aircraft," *IEEE Trans. on Automatic Control*, vol. AC-22, no. 5, pp. 753–757, 1977.

[9] J. Roskam, *Airplane Flight dynamics and Automatic Flight Controls: Part II*. Kansas City: Roskam Aviation and Engineering Corp., 1979.

[10] J. Roskam, *Airplane Design Part VII: Determination of Stability, Control and Performance Charactersitics–FAR and Military Requirements*. Route 4, Box 274, Ottawa, Kansas, 66067: Roskam Aviation and Engineering Corporation, 1985.

Limiting Performance of Optimal Linear Filters*

J.H. Braslavsky
julio@seidel.ece.ucsb.edu

M.M. Seron
seron@seidel.ece.ucsb.edu

D.Q. Mayne
mayne@ece.ucdavis.edu

P.V. Kokotović
petar@ece.ucsb.edu

Center for Control Engineering and Computations
University of California
Santa Barbara, CA 93106-9560

Abstract

We study the lowest achievable mean square estimation error in two limiting optimal linear filtering problems. First, when the intensity of the process noise tends to zero, the lowest achievable mean square estimation error is a function of the unstable poles of the system. Second, when the intensity of the measurement noise tends to zero, the lowest achievable mean square estimation error is a function of the nonminimum phase zeros of the system. We link these results with Bode integral characterizations of performance limitations in linear filtering.

Keywords: Optimal Linear Filtering, Performance Limits, Nonminimum Phase Zeros, Unstable Poles, Bode Integrals, Singular Perturbations.

1 Introduction

The analysis of feedback limitations using analytic function theory was initiated by Bode in the 1940's and is of continuing interest [1, 6, 7, 14, 4]. In recent studies of performance limitations in filtering, Bode and Poisson integrals were established for functions representing sensitivity of the estimation error to process and measurement noise [8, 16].

An alternative approach is to study the best achievable performance of optimal control when the cost of control tends to zero (cheap control) [3, 9, 13, 5, 18, 15]. This approach was recently linked to Bode integrals and extended to nonlinear systems [17].

In this paper we derive performance limits in optimal filtering and establish their links with Bode integrals. By letting the intensity of either the process noise or the measurement noise tend to zero (as in [2]), we show that these performance limits are imposed by the unstable poles and nonminimum-phase (NMP) zeros of the transfer function from the process noise input to the measurements.

For vanishing process noise, the lowest achievable mean square (MS) estimation error must be nonzero if the system is unstable. In particular, when estimating the specific combination of

---

*Research supported in part by the National Science Foundation under grant ECS-9528370 and the Air Force Office of Scientific Research under grant F49620-95-1-0409.

states that form the measurements, we show that this error is an explicit function of the unstable poles, further characterized with a Bode integral.

For vanishing measurement noise, the lowest achievable MS estimation error must be nonzero if the system is NMP [13]. We show that this error corresponds to that achieved by a reduced filtering problem for the unstable zero-dynamics of the system. In particular, for the problem of estimating a slowly-varying input to the system, the lowest achievable MS error is an explicit function of the NMP zeros, also characterized with a Bode integral.

Our filtering problem is the optimal MS estimation [10] of the state x of the linear time-invariant system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bw(t), \quad x \in \mathbb{R}^n, w \in \mathbb{R}^m, \\ y(t) &= Cx(t) + v(t), \quad y \in \mathbb{R}^p, \end{aligned} \tag{1}$$

where B and C are full rank matrices and w and v are uncorrelated zero-mean Gaussian white noises with intensities Σ_w and Σ_v, that is,

$$\begin{aligned} E\{w(t)\} &= 0, \qquad E\{w(t)w^T(\tau)\} = \Sigma_w \,\delta(t-\tau), \\ E\{v(t)\} &= 0, \qquad E\{v(t)v^T(\tau)\} = \Sigma_v \,\delta(t-\tau). \end{aligned}$$

The matrices Σ_w and Σ_v are assumed constant and positive definite. The initial state $x(0)$ is a Gaussian random vector with mean $\bar{x}_0$ and covariance Σ_0.

The steady-state optimal filter for the system (1) is

$$\dot{\hat{x}}(t) = A\hat{x}(t) + QC^T\Sigma_v^{-1}[y(t) - C\hat{x}(t)], \qquad \hat{x}(0) = \bar{x}_0, \tag{2}$$

where Q is a positive semidefinite solution of the algebraic Riccati equation (ARE)

$$AQ + QA^T + B\Sigma_w B^T - QC^T\Sigma_v^{-1}CQ = 0. \tag{3}$$

The matrix Q, which is unique if (1) is stabilizable and detectable, is the asymptotic variance of the estimation error, that is

$$\lim_{t\to\infty} E\{[x(t) - \hat{x}(t)][x(t) - \hat{x}(t)]^T\} = Q.$$

The optimal filter (2) minimizes the MS estimation error

$$\lim_{t\to\infty} E\{[x(t) - \hat{x}(t)]^T[x(t) - \hat{x}(t)]\} = \operatorname{trace} Q.$$

2 Performance Limits Under Small Process Noise

2.1 Limiting Properties

Assumption 1.

(i) The system (1) is stabilizable and detectable.

(ii) The matrix A in (1) has no eigenvalues on the imaginary axis.

(iii) The intensities of the noises w and v are

$$\Sigma_w = \varepsilon^2 I, \quad \text{and} \quad \Sigma_v = I, \tag{4}$$

where $\varepsilon > 0$ is small. ○

Conditions (i) and (ii) are necessary and sufficient for the filter (2) to be asymptotically stable in the limit as w vanishes. Because of (ii), a modal decomposition performed on (1) brings it to the form (Figure 1)

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 + B_1 w, \\ \dot{x}_2 &= A_2 x_2 + B_2 w, \\ y &= C_1 x_1 + C_2 x_2 + \nu, \end{aligned} \tag{5}$$

where A_1 is Hurwitz and A_2 is *antistable* (that is $-A_2$ is Hurwitz). Condition (i) implies that (A_2, B_2, C_2) is controllable and observable. In the form (5), the eigenvalues of A_2 are the unstable poles of the system, that is, those of the transfer function from w to y.

We denote by $(\hat{x}_1, \hat{x}_2)$ the estimate of the state of (5) and by $(\tilde{x}_1, \tilde{x}_2) = (x_1 - \hat{x}_1, x_2 - \hat{x}_2)$ the estimation error.

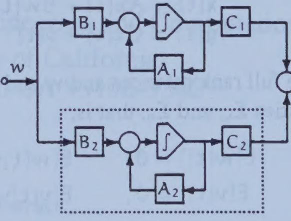


Figure 1: Modal decomposition of system (5)

Lemma 2.1 (Error Variance under No Process Noise).
Under Assumption 1, the variance of the estimation error achieved by the optimal filter for (5) satisfies

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} E\left\{ \begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \end{bmatrix} \begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \end{bmatrix}^T \right\} = \begin{bmatrix} 0 & 0 \\ 0 & \bar{Q}_2 \end{bmatrix}, \tag{6}$$

and the lowest achievable MS estimation error is

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} E\left\{ \begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \end{bmatrix}^T \begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \end{bmatrix} \right\} = \operatorname{trace} \bar{Q}_2, \tag{7}$$

where $\bar{Q}_2$ is the positive definite solution of the ARE

$$A_2 \bar{Q}_2 + \bar{Q}_2 A_2^T = \bar{Q}_2 C_2^T C_2 \bar{Q}_2. \tag{8}$$

○

The matrix $\bar{Q}_2$ in (6) is the variance that would result from the filtering problem for

$$\begin{aligned} \dot{x}_2 &= A_2 x_2, \\ y_2 &= C_2 x_2 + \nu, \end{aligned}$$

the unstable subsystem in (5) (Figure 1). Hence, only the unstable modes of the system contribute to the MS estimation error (trace $\bar{Q}_2$) when the process noise vanishes.

The optimal filter (2) takes the form

$$\begin{aligned} \dot{\hat{x}}_1 &= A_1 \hat{x}_1 + \varepsilon^2 (\bar{Q}_1 C_1^T + \bar{Q}_3 C_2^T)(y - C_1 \hat{x}_1 - C_2 \hat{x}_2), \\ \dot{\hat{x}}_2 &= A_2 \hat{x}_2 + (\varepsilon^2 \bar{Q}_3^T C_1^T + \bar{Q}_2 C_2^T)(y - C_1 \hat{x}_1 - C_2 \hat{x}_2) \end{aligned} \tag{9}$$

and, in the limit as $\varepsilon \to 0$,

$$\begin{aligned} \dot{\hat{x}}_1 &= A_1\hat{x}_1 \,, \\ \dot{\hat{x}}_2 &= -\bar{Q}_2 A_2^T \bar{Q}_2^{-1}\hat{x}_2 + \bar{Q}_2 C_2^T (y - C_1\hat{x}_1) \,, \end{aligned} \tag{10}$$

where we have used $A_2 - \bar{Q}_2 C_2^T C_2 = -\bar{Q}_2 A_2^T \bar{Q}_2^{-1}$, which follows from (8). We see from (10) that the eigenvalues of the optimal filter converge, as $\varepsilon \to 0$, to the stable eigenvalues of (5) and to the *mirror image* (with respect to the imaginary axis) of the unstable poles of the system [13].

As the process noise vanishes, the filtering problem becomes that of estimating the noise-free state of the system (5) using noisy measurements. Expression (10) shows that the filter for the stable subsystem $\dot{x}_1 = A_1 x_1$ is disconnected from the measurements. However, the filter for the unstable subsystem $\dot{x}_2 = A_2 x_2$ must continue to be driven by the noisy measurements due to the requirement of stability of the filter imposed by the optimization problem. Under this requirement, the lowest MS estimation error is achieved when the eigenvalues of the filter are located at the mirror image of the unstable poles of the system (5).

2.2 A Bode Integral for the Estimation of the Noise-Free Output

Define $\psi = C_1 x_1 + C_2 x_2$ as the noise-free "output" of system (5) (Figure 1), and let $\hat{\psi}$ denote the estimate of ψ and $\tilde{\psi} = \psi - \hat{\psi}$ the estimation error.

Theorem 2.2 (Performance Limits Imposed by Unstable Poles). For the system (5) under Assumption 1, the lowest achievable MS error in the estimation of $\psi = C_1 x_1 + C_2 x_2$ satisfies

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} E\left\{\tilde{\psi}(t)^T \tilde{\psi}(t)\right\} = 2 \sum_{i=1}^{n_2} \beta_i \,, \tag{11}$$

where $\beta_1, \ldots, \beta_{n_2}$ are the system unstable poles. ○

Theorem 2.2 shows that the lowest achievable MS estimation error in the problem of estimating ψ is precisely quantified as twice the sum of the system unstable poles.

This theorem is dual to the feedback control result that characterizes the right hand side of (11) as the optimal cost achieved by the *minimum energy control* transferring the system (5) to rest from the initial condition originating from a unit impulse at the input [17]. The optimal cost for this minimum energy problem is equal to a feedback invariant quantity that involves the well-known Bode integral for the sensitivity function of a unitary feedback loop.

Bode integrals have also been obtained for transfer functions that represent sensitivity properties of filtering systems [8, 16]. The following theorem links one of these integrals with the lowest achievable MS error in the estimation of the noise-free output ψ of the system (5).

Theorem 2.3 (Bode Invariant for the Estimation of the Noise-Free Output). Let G be the transfer function from w to y in (5), and F be the transfer function of any stable, strictly proper filter mapping the system output y to the estimate $\hat{\psi} = C_1\hat{x}_1 + C_2\hat{x}_2$. Assume that F is such that the estimation error transfer function $(1 - F)G$ is stable and minimum phase. Then

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \log|1 - F(j\omega)| \, d\omega + \lim_{s \to \infty} sF(s) = \lim_{t \to \infty} E\left\{\tilde{\psi}^2(t)\right\} \,, \tag{12}$$

where $\lim_{t\to\infty} E\left\{\tilde{\psi}^2(t)\right\}$ is the lowest achievable MS error in the estimation of $\psi = C_1 x_1 + C_2 x_2$ in (5), (4). ○

The transfer function $1 - F$ from ψ to $\tilde{\psi}$ (Figure 2) quantifies the performance of the filter: the smaller this magnitude is over the frequency range where the spectrum of ψ is concentrated, the better the estimate $\hat{\psi}$ "tracks" the signal ψ. In this sense, $1 - F$ is a *filtering sensitivity function* dual to the sensitivity function S of the unitary feedback control loop [16, § 7.2].

Any filter F satisfying the conditions of Theorem 2.3 must satisfy (12), which equals twice the sum of the unstable poles of the system. We may thus view the LHS of (12) as a filtering *sensitivity invariant* in the sense that its terms must be traded-off against each other to add to a quantity fixed by the system and *independent of the filter parameters*. Theorem 2.3 states that this filtering invariant is the MS estimation error achieved by the optimal filter when there is no process noise.

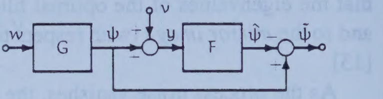


Figure 2: Estimation of the noise-free output

When F is the *optimal* filter that estimates $\hat{\psi} = C_1\hat{x}_1 + C_2\hat{x}_2$, from (10) we have that

$$\begin{aligned} F(s) &= C_2(sI + \bar{Q}_2A_2^T\bar{Q}_2^{-1})^{-1}\bar{Q}_2C_2^T, \\ &= C_2\bar{Q}_2(sI + A_2^T)^{-1}C_2^T, \end{aligned} \tag{13}$$

has relative-degree one ($C_2\bar{Q}_2C_2^T \neq 0$), and its poles are the mirror image of the unstable poles of the system. Also, $1 - F$ is a relative-degree-zero transfer function with zeros at the unstable poles of the system. Hence $1 - F$ is *all-pass*, that is, $|1 - F(j\omega)| = 1$, and we conclude that $\frac{1}{\pi}\int_{-\infty}^{\infty} \log|1 - F(j\omega)|\, d\omega = 0$, and therefore $\lim_{s\to\infty} sF(s) = 2\sum_{i=1}^{n_2}\beta_i$.

3 Performance Limits Under Small Measurement Noise

3.1 Limiting Properties

Assumption 2.

(i) The system (1) is stabilizable and detectable.

(ii) The system (1) has relative-degree one: rank $CB = m$.

(iii) All the zeros of $C(sI - A)^{-1}B$ are nonminimum phase.

(iv) The intensities of the noises w and ν are

$$\Sigma_w = I, \quad \text{and} \quad \Sigma_\nu = \varepsilon^2 I, \tag{14}$$

where $\varepsilon > 0$ is small. ○

Assumptions (ii) and (iii), which represent the case with the maximum number of NMP zeros, are made only for simplicity of exposition and may be removed.

Let M_0 be a basis of the left null-space of the input matrix B (that is, $M_0B = 0$). Then the change of coordinates $\xi = Cx$ and $z = M_0x$ transforms (1) to the form

$$\begin{aligned} \dot{\xi} &= A_1\xi + C_0z + B_1w, \\ \dot{z} &= A_0z + B_0\xi, \\ y &= \xi + \nu. \end{aligned} \tag{15}$$

This form is of interest because it displays the zeros of $C(sI - A)^{-1}B$ (the transfer function from w to y in (1)) as the eigenvalues of A_0. When $\nu = 0$ (so $y = \xi$), (15) can be seen to consist of two systems in feedback interconnection: the zero dynamics subsystem described by $\dot{z} = A_0 z + B_0 y$, with input y and output $C_0 z$, and the output subsystem $\dot{y} = A_1 y + C_0 z + B_1 w$, with input $C_0 z + B_1 w$ and output y (Figure 3). The triple (A_0, B_0, C_0) is controllable and observable because (15) is stabilizable and detectable and A_0 is antistable by (iii), Assumption 2.

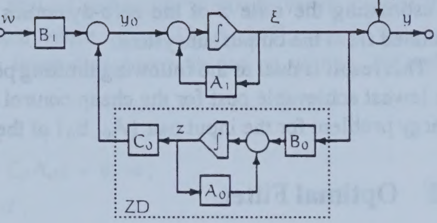


Figure 3: Output/zero-dynamics interaction in the system (15)

We denote by $(\hat{\xi}, \hat{z})$ the estimate of the state of (15) and by $(\tilde{\xi}, \tilde{z}) = (\xi - \hat{\xi}, z - \hat{z})$ the estimation error.

Lemma 3.1 (Error Variance under No Measurement Noise). Under Assumption 2, the variance of the estimation error achieved by the optimal filter for (15) satisfies

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} E \left\{ \begin{bmatrix} \tilde{\xi}(t) \\ \tilde{z}(t) \end{bmatrix} \begin{bmatrix} \tilde{\xi}(t) \\ \tilde{z}(t) \end{bmatrix}^T \right\} = \begin{bmatrix} 0 & 0 \\ 0 & \bar{Q}_0 \end{bmatrix}, \tag{16}$$

and the lowest achievable MS estimation error is

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} E \left\{ \begin{bmatrix} \tilde{\xi}(t) \\ \tilde{z}(t) \end{bmatrix}^T \begin{bmatrix} \tilde{\xi}(t) \\ \tilde{z}(t) \end{bmatrix} \right\} = \operatorname{trace} \bar{Q}_0, \tag{17}$$

where $\bar{Q}_0$ is the positive definite solution of

$$A_0 \bar{Q}_0 + \bar{Q}_0 A_0^T = \bar{Q}_0 C_0^T (B_1 B_1^T)^{-1} C_0 \bar{Q}_0. \tag{18}$$

○

If A_0 were Hurwitz then $\bar{Q}_0 = 0$ would be the nonnegative definite solution of (18), and the error variance achieved by the optimal filter as $\varepsilon \to 0$ would be zero. In our case A_0 is antistable, and hence the error variance is positive definite.

In the limit when the measurement noise vanishes, $\nu = 0$, the state $\xi = y$ is exactly measured. Therefore, in the output system $\dot{\xi} - A_1 \xi = C_0 z + B_1 w$ we can treat $y_0 = \dot{\xi} - A_1 \xi$ as the noisy measurement $y_0 = C_0 z + B_1 w$ of the state of the unstable zero dynamics subsystem (Figure 3). This is equivalent to the filtering problem for the system

$$\begin{aligned} \dot{z} &= A_0 z, \\ y_0 &= C_0 z + B_1 w. \end{aligned} \tag{19}$$

The variance of the estimation error corresponding to this filtering problem is $\bar{Q}_0$, the positive definite solution of (18). We see that the limiting estimation problem (19) is characterized by the output pair (A_0, C_0) of the zero-dynamics subsystem with the driving term $B_0 \xi$ treated as a known input.

Hence, only the unstable zero dynamics of the system contribute to the MS estimation error (trace $\bar{Q}_0$) when the measurement noise vanishes. In other words, the problem reduces to that of estimating the state z of the zero-dynamics subsystem given the noisy measurement y_0 obtained from the output subsystem.

This result is dual to the following limiting property for the *cheap control problem* for (15): the lowest achievable cost for the cheap control is the optimal cost achieved by the minimum energy problem for the input pair (A_0, B_0) of the zero-dynamics subsystem [17].

3.2 Optimal Filter

By dualizing the cheap control results, Kwakernaak and Sivan [13] showed that, in the limit as $\varepsilon \longrightarrow 0$, the eigenvalues of the optimal filter which do not converge to infinity converge to the system minimum phase zeros and to the mirror image of the NMP zeros. We recover this result using singular perturbation analysis.

Note first that, since $\Sigma_\nu = \varepsilon^2 I$, we can express y in (15) as $y = \xi + \varepsilon \nu_1$, where ν_1 is a zero-mean Gaussian white noise with covariance $E\{\nu_1(t)\nu_1{}^T(\tau)\} = I\,\delta(t-\tau)$. The optimal filter (2) takes the form

$$\begin{aligned} \dot{\hat{\xi}} &= A_1\hat{\xi} + C_0\hat{z} + \frac{1}{\varepsilon}\left[(B_1B_1^T)^{1/2} + \varepsilon\bar{\bar{Q}}_1 + O(\varepsilon^2)\right](\xi + \varepsilon\nu_1 - \hat{\xi}), \\ \dot{\hat{z}} &= B_0\hat{\xi} + A_0\hat{z} + \frac{1}{\varepsilon}\left[\bar{Q}_0C_0^T(B_1B_1^T)^{-1/2} + \varepsilon\bar{\bar{Q}}_2 + O(\varepsilon^2)\right](\xi + \varepsilon\nu_1 - \hat{\xi}). \end{aligned} \tag{20}$$

Since $B_1B_1^T$ is nonsingular, the filter (20) is a singularly perturbed system. To identify its fast and slow subsystems we perform the change of variable $\eta = \hat{z} - \bar{Q}_0C_0^T(B_1B_1^T)^{-1}\hat{\xi}$, which brings (20) into the standard singular perturbation form

$$\begin{aligned} \varepsilon\dot{\hat{\xi}} &= (B_1B_1^T)^{1/2}(\xi - \hat{\xi} + \varepsilon\nu_1) + O(\varepsilon), \\ \dot{\eta} &= B_\eta\hat{\xi} - \bar{Q}_0A_0^T\bar{Q}_0^{-1}\eta + K_\eta(\xi - \hat{\xi} + \varepsilon\nu_1) + O(\varepsilon), \end{aligned} \tag{21}$$

where $B_\eta = B_0 - \bar{Q}_0[C_0^T(B_1B_1^T)^{-1}A_1 + A_0^TC_0^T(B_1B_1^T)^{-1}]$, $K_\eta = \bar{\bar{Q}}_2 - \bar{Q}_0C_0^T(B_1B_1^T)^{-1}\bar{\bar{Q}}_1$, and where we have used $A_0 - \bar{Q}_0C_0^T(B_1B_1^T)^{-1}C_0 = -\bar{Q}_0A_0^T\bar{Q}_0^{-1}$, which follows from (18). Khalil and Gajic [11, Theorem 1] have shown that the system

$$\begin{aligned} \varepsilon\dot{\hat{\xi}}_0 &= (B_1B_1^T)^{1/2}(\xi - \hat{\xi}_0), \\ \dot{\eta}_0 &= B_\eta\hat{\xi}_0 - \bar{Q}_0A_0^T\bar{Q}_0^{-1}\eta_0 + K_\eta(\xi - \hat{\xi}_0) \end{aligned} \tag{22}$$

is an $O(\varepsilon)$ approximation of (21) in the sense that the variance of $\hat{\xi} - \hat{\xi}_0$ is $O(\varepsilon)$ and the variance of $\eta - \eta_0$ is $O(\varepsilon^2)$. Since (22) is noise-free and $(B_1B_1^T)^{1/2}$ is positive definite, it is well known [12] that, as $\varepsilon \longrightarrow 0$, the fast variable $\hat{\xi}_0$ experiences an initial *boundary-layer transient* in the fast time scale t/ε towards its *quasi-steady state* $\hat{\xi}_0 = \xi$. The quasi-steady state of the slow variable η_0, obtained by substitution of $\hat{\xi}_0 = \xi$ in the η_0-subsystem of (22), is $\dot{\eta}_0 = -\bar{Q}_0A_0^T\bar{Q}_0^{-1}\eta_0 + B_\eta\xi$. Thus, as the noise $\varepsilon\nu_1$ vanishes, the filter (20) reduces to

$$\begin{aligned} \hat{\xi} &= \xi, \\ \dot{\eta} &= -\bar{Q}_0A_0^T\bar{Q}_0^{-1}\eta + B_\eta\xi, \\ \hat{z} &= \eta + \bar{Q}_0C_0^T(B_1B_1^T)^{-1}\xi. \end{aligned} \tag{23}$$

The eigenvalues of $-\bar{Q}_0A_0^T\bar{Q}_0^{-1}$ are the *mirror image* of the eigenvalues of A_0, the NMP zeros of (15). It follows that, in the limit as $\varepsilon \longrightarrow 0$, $n - m$ eigenvalues of the optimal filter are the mirror image of the NMP zeros of the system and the remaining m eigenvalues lie at infinity.

3.3 Estimation of a Random-Walk Input

A filtering problem of practical relevance is that of estimating a slowly-varying input disturbance, which may be modeled as integrated white noise [13, p. 347], that is "random walk". The system (15) with a random-walk input disturbance ψ can be written, by defining $\zeta = C_0 z + B_1 \psi$, as

$$\begin{aligned} \dot{\xi} &= A_1 \xi + \zeta , \\ \dot{\zeta} &= C_0 B_0 \xi + C_0 A_0 z + B_1 w , \\ \dot{z} &= B_0 \xi + A_0 z , \\ y &= \xi + \nu . \end{aligned} \tag{24}$$

where w and ν are zero-mean Gaussian white noises of intensities given by (14).

Let $\hat{\zeta}$ and $\hat{\psi} = B_1^{-1}(\hat{\zeta} - C_0 \hat{z})$ be the estimates of the state ζ of (24) and of the disturbance ψ, and let $\tilde{\zeta} = \zeta - \hat{\zeta}$ and $\tilde{\psi} = \psi - \hat{\psi}$ be the estimation errors.

Theorem 3.2 (Performance Limits Imposed by NMP Zeros). Assume that the system (24) is stabilizable and detectable, with B_1 invertible and A_0 antistable. Then the filtering problem for (24), (14) has the following limiting properties:

(i) The variance of the optimal estimation error satisfies

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} E \left\{ \begin{bmatrix} \tilde{\xi}(t) \\ \tilde{\zeta}(t) \\ \tilde{z}(t) \end{bmatrix} \begin{bmatrix} \tilde{\xi}(t) \\ \tilde{\zeta}(t) \\ \tilde{z}(t) \end{bmatrix}^T \right\} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_0 \end{bmatrix} , \tag{25}$$

where $\bar{Q}_0$ is the positive definite solution of

$$A_0 \bar{Q}_0 + \bar{Q}_0 A_0^T = \bar{Q}_0 A_0^T C_0^T (B_1 B_1^T)^{-1} C_0 A_0 \bar{Q}_0 . \tag{26}$$

(ii) The lowest achievable MS error in the estimation of the input disturbance ψ is

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} E \left\{ \tilde{\psi}(t)^T \tilde{\psi}(t) \right\} = 2 \sum_{i=1}^{n-m} \frac{1}{\alpha_i} , \tag{27}$$

where $\alpha_1, \ldots, \alpha_{n-m}$ are the NMP zeros of the system. ○

Theorem 3.2 states that there is a nonzero MS error in the estimation of a random-walk input to a NMP system. This error is an explicit function of the NMP zeros, independent of the system realization, and which will be larger when the NMP zeros are closer to the origin. This result is dual to Theorem 3 in [15], which states that the lowest achievable cost for the cheap control problem of regulating the output y of (15) to a constant setpoint r is $\frac{1}{2} r^T H r$, where H is an Hermitian matrix whose trace equals (27).

A singular perturbation analysis similar to that of Section 3.2 shows that the optimal filter for (24) estimating $\psi = B_1^{-1}(\zeta - C_0 z)$ is

$$\begin{aligned} \hat{\xi} &= \xi , \\ \hat{\zeta} &= \dot{\xi} - A_1 \xi , \\ \dot{\gamma} &= (I - KC_0) A_0 \gamma + (I - KC_0) B_0 \xi + (I - KC_0) A_0 K (\dot{\xi} - A_1 \xi) , \\ \hat{z} &= \gamma + K \hat{\zeta} , \\ \hat{\psi} &= B_1^{-1} \left[(I - C_0 K)(\dot{\xi} - A_1 \xi) - C_0 \gamma \right] . \end{aligned} \tag{28}$$

Again, the eigenvalues of the filter (28) are the mirror image of the NMP zeros of the system, but the filter is driven by both the system output $y = \xi$ and its first derivative $\dot{y} = \dot{\xi}$. Since it has a feedthrough from $\dot{y}$ to the estimates, its relative degree is minus one.

3.4 A Bode Integral for the Estimation of a Random-Walk Input

Theorem 3.3 (Bode Invariant for the Estimation of a Random-Walk Input). Let G be the transfer function from ψ to y in (24), and F be the transfer function of any stable filter mapping the system output y to the estimate $\hat{\psi}$. Assume that F is such that

(i) the estimation error transfer function $1 - FG$ is stable and proper,

(ii) the NMP zeros of FG are equal to the NMP zeros of G.

Then

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \log|F(j\omega)G(j\omega)| \frac{d\omega}{\omega^2} - \lim_{s\to 0}\frac{d(FG)}{ds} = \lim_{t\to\infty} E\left\{\tilde{\psi}^2(t)\right\}, \tag{29}$$

where $\lim_{t\to\infty} E\left\{\tilde{\psi}^2(t)\right\}$ is the lowest achievable MS error in the estimation of the random-walk input ψ in (24), (14). ○

The transfer function FG from ψ to $\hat{\psi}$ (Figure 4) is a *filtering complementary sensitivity function* that quantifies the filter performance [16, § 7.2].

The integral in (29) measures the filter performance since its value gives the difference between areas of complementary sensitivity amplification and attenuation. The second term on the LHS of (29) equals the steady-state error $\tilde{\psi}$ that would result if ψ were a ramp signal. For *any* filter F satisfying the conditions of Theorem 3.3, (29) is a fixed quantity equal to twice the sum of the reciprocal of the NMP zeros of the system. Thus, the LHS of (29) is a filtering *complementary sensitivity invariant* in the sense that its terms must be traded-off against each other to add to this fixed quantity. Theorem 3.3 states that this invariant is the MS estimation error achieved by the optimal filter for this problem under no measurement noise.

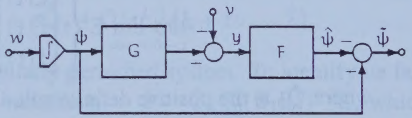


Figure 4: Estimation of a random-walk input disturbance

When F is the optimal filter the filtering complementary sensitivity FG is all-pass, which implies that $\frac{1}{\pi}\int_{-\infty}^{\infty} \log|F(j\omega)G(j\omega)| \frac{d\omega}{\omega^2} = 0$, and therefore $-\lim_{s\to 0}\frac{d(FG)}{ds} = 2\sum_{i=1}^{n-m}\frac{1}{\alpha_i}$. The absence of measurement noise allows the optimal filter to "invert" the system G (with its NMP zeros mirrored with respect to the imaginary axis) by using derivatives of the system output. Thus, in the case of zero measurement noise, an "ideal" complementary sensitivity FG having magnitude equal to one at all frequencies is achievable.

References

[1] H.W. Bode. *Network Analysis and Feedback Amplifier Design.* D. van Nostrand, New York, 1945.

[2] A.E. Bryson and D.E. Johansen. Linear filtering for time-varying systems using measurements containing colored noise. *IEEE Trans. on Automatic Control*, 10:4–10, January 1965.

[3] S.S.L. Chang. *Synthesis of optimum control systems.* McGraw-Hill, 1961.

[4] J. Chen. Sensitivity integral relations and design trade-offs in linear multivariable feedback systems. *IEEE Trans. on Automatic Control*, 40(10):1700–1716, October 1995.

[5] B. Francis. The optimal linear-quadratic time-invariant regulator with cheap control. *IEEE Trans. on Automatic Control*, 24:616–621, 1979.

[6] B.A. Francis and G. Zames. On H_∞-optimal sensitivity theory for SISO feedback systems. *IEEE Trans. on Automatic Control*, 29(1):pp. 9–16, 1984.

[7] J.S. Freudenberg and D.P. Looze. Right half plane poles and zeros and design tradeoffs in feedback systems. *IEEE Trans. on Automatic Control*, 30(6):555–565, June 1985.

[8] G.C. Goodwin, D.Q. Mayne, and J. Shim. Trade-offs in linear filter design. *Automatica*, 31(10):1367–1376, Oct. 1995.

[9] R.E. Kalman. When is a linear control system optimal? *Trans. ASME. J. Basic Engrg.*, 86:51–60, 1964.

[10] R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME. J. Basic Engrg.*, 83-D:95–107, 1961.

[11] H.K. Khalil and Z. Gajic. Near-optimum regulators for stochastic linear singularly perturbed systems. *IEEE Trans. on Automatic Control*, 29:531–541, June 1984.

[12] P.V. Kokotović, H.K. Khalil, and J. O'Reilly. *Singular perturbation methods in control: Analisys and Design.* Academmic Press, 1986.

[13] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, 1972.

[14] R.H. Middleton. Trade-offs in linear control systems design. *Automatica*, 27(2):281–292, March 1991.

[15] L. Qiu and E.J. Davison. Performance limitations of non-minimum phase systems in the servomechanism problem. *Automatica*, 29(2):337–349, 1993.

[16] M.M. Seron, J.H. Braslavsky, and G.C. Goodwin. *Fundamental Limitations in Filtering and Control.* CCES. Springer, 1997.

[17] M.M. Seron, J.H. Braslavsky, and P.V. Kokotović. Feedback limitations in nonlinear systems: From bode integrals to cheap control. In *Proc. of the 35th IEEE CDC*, San Diego, USA, 1997. To appear.

[18] U. Shaked. Singular and cheap optimal control: the minimum and nonminimum phase cases. Technical Report TWISK 181, National Research Institute for Mathematical Sciences, Pretoria, Republic of South Africa, 1980.

A Design Tradeoff for Single Input Two Output Feedback Systems*

A.R. Woodyatt[†] J.S. Freudenberg[‡]and R.H. Middleton[§]

Department of Electrical and Computer Engineering
The University of Newcastle, NSW 2308 Australia

Abstract

Consider a linear feedback system for which the plant has a single input and two outputs. Suppose that the direction of the plant changes significantly with frequency. We show that such plants are subject to inherent design limitations that become important if the change in direction occurs within the desired closed loop bandwidth. The result is stated in terms of an integral relation that must be satisfied by the output complementary sensitivity function.

1 Introduction

The theory of inherent design limitations for Single Input, Single Output (SISO) linear time-invariant feedback systems is well established. Recent years have seen this theory extended to sampled-data, nonlinear, and Multiple Input, Multiple Output (MIMO) feedback systems. A survey of results and references are found in [3].

Despite progress, the theory of design tradeoffs for MIMO feedback systems is less developed than that for SISO systems. In this paper, we suppose focus on a special case, and suppose that the plant has one control actuator and two measured outputs. We refer to such a plant as Single Input, Two Output (SITO). This class of feedback systems was studied in [2], wherein it was argued that Two Input, Two Output (TITO) feedback systems with an actuator bandwidth limitation may be usefully viewed as SITO systems for the purpose of controlling transient response.

The concepts of plant and controller direction and alignment were used in [2] to describe design limitations and tradeoffs for a SITO feedback system at a fixed frequency. Such tradeoffs are termed "algebraic", and are due to the structure associated with a feedback loop. More subtle tradeoffs are those that occur between system properties in different frequency ranges; such tradeoffs are termed "analytic". Examples are the various sensitivity and complementary sensitivity integrals surveyed in [3]. *Our purpose in the present paper is to show that there exist analytic tradeoffs for SITO systems that have no counterpart in the SISO case.* We shall see that these tradeoffs are present

---

*Work supported by the Australian Research Council, U. S. National Science Foundation, and Ford Motor Co.

[†] `woody@ee.newcastle.edu.au`

[‡] J.S. Freudenberg is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122 USA, `jfr@eecs.umich.edu`

[§] `rick@ee.newcastle.edu.au`

whenever the direction of the plant changes significantly with frequency over the desired closed loop bandwidth.

The remainder of this paper is outlined as follows. Section 2 introduces our notation. In Section 3 we review the concepts of plant and controller direction and alignment. These were used in [2] to describe the relation between feedback properties at the plant input and output. We also present a new result relating alignment to the extent of closed loop interactions. These results suggest that a good design will have the controller "well-aligned" with the plant at frequencies within the desired closed loop bandwidth. In Section 4 we study the extreme case in which the controller is perfectly aligned with the plant at all frequencies. We show that if the direction of the plant changes with frequency, then the resulting feedback system will possess nonminimum phase zeros and associated bandwidth limitations. Because perfect alignment at all frequencies is unnecessary to achieve a satisfactory closed loop response, we proceed to explore design limitations when the alignment constraint is relaxed. In Section 5 we describe a coordinate transformation that allows us to easily describe the extent to which plant direction varies from that at DC. In Section 6 we state the major result of the paper, which is an integral relation that must be satisfied by the output complementary sensitivity function. This result is illustrated in Section 7. Conclusions are stated in Section 8.

2 Preliminaries

Given a vector $v \in \mathbf{C}^n$, let $\|v\|$ denote the Euclidean vector norm. Given a matrix $M \in \mathbf{C}^{m\times n}$, denote the range, or column space of M as $\mathcal{R}(M)$, the orthogonal complement of $\mathcal{R}(M)$ by $\mathcal{R}^{\perp}(M)$, the nullspace of M by $\mathcal{N}(M)$, the rowspace of M by $\mathcal{R}_{row}(M)$, and the left nullspace of M by $\mathcal{N}_{left}(M)$. Denote the Euclidean matrix norm of M by $\|M\|$. We denote the open and closed right and left halves of the complex plane by ORHP, CRHP, OLHP, and CLHP, respectively. Let $G(s)$ be a scalar rational function, and suppose that $G(z) = 0$ for $z \in$ ORHP. Then we say that $G(s)$ is non-minimum phase (NMP), and refer to z as a NMP zero. If $M(s)$ is stable and proper, denote the H_∞ norm of $M(s)$ by $\|M\|_\infty \triangleq \sup_{\omega\in\mathbf{R}} \|M(j\omega)\|$.

Consider a linear time-invariant SITO plant, controlled by a linear time-invariant controller. Let the transfer functions of the plant and controller be given by $P(s) = \begin{bmatrix} p_1(s) & p_2(s)\end{bmatrix}^T$ and $C(s) = \begin{bmatrix} c_1(s) & c_2(s)\end{bmatrix}$, respectively, where $p_1(s)$, $p_2(s)$, $c_1(s)$, and $c_2(s)$ are rational and proper.

Define the *input and output open loop transfer functions*, $L_I(s) \triangleq C(s)P(s)$ and $L_O(s) \triangleq P(s)C(s)$, the *sensitivity functions*, $S_I(s) \triangleq (1 + L_I(s))^{-1}$ and $S_O(s) \triangleq (I + L_O(s))^{-1}$, and the *complementary sensitivity functions*, $T_I(s) \triangleq L_I(s)\,(1 + L_I(s))^{-1}$ and $T_O(s) \triangleq L_O(s)\,(I + L_O(s))^{-1}$. The closed loop transfer functions each describe important feedback properties of the closed loop system. For example, the response of the system output to exogenous inputs is given by

$$Y(s) = S_O(s)D_O(s) + P(s)S_I(s)D_I(s) - T_O(s)N(s) + T_O(s)R(s). \qquad (2.1)$$

Finally, we say that the feedback system is *stable* if the plant and controller each have no hidden modes in the CRHP, and if the transfer functions $S_O(s)$, $S_I(s)$, $C(s)S_O(s)$, and $S_O(s)P(s)$ have no poles in the CRHP.

3 Plant/Controller Direction and Alignment

We now review the concepts of plant and controller *direction*, and plant/controller *alignment*. The following definitions are taken from [2].

Definition 3.1 (Direction and Alignment Angle) *Identify the* directions *of the plant and controller at frequency ω with $\mathcal{R}(P(j\omega))$ and $\mathcal{R}_{row}(C(j\omega))$, respectively*[1]. *Define the* plant/controller alignment angle *by*

$$\cos\phi(j\omega) = \frac{|C(j\omega)P(j\omega)|}{\|C(j\omega)\|\|P(j\omega)\|}. \tag{3.1}$$

□

It is straightforward to verify that the plant and controller directions and alignment angle depend only upon the *plant and controller ratios* $P_{rat}(j\omega) \triangleq p_2(j\omega)/p_1(j\omega)$ and $C_{rat}(j\omega) \triangleq c_2(j\omega)/c_1(j\omega)$. As discussed in [2], the alignment angle is a measure of how closely the direction of the controller matches that of the plant.

Definition 3.2 (Perfect Alignment and Complete Misalignment)

(i). If $\phi(j\omega) = 0°$ then we say that the plant and controller are perfectly aligned. *Perfect alignment requires that the gain and phase of the controller ratio satisfy $|C_{rat}(j\omega)| = |P_{rat}(j\omega)|$ and $\angle C_{rat}(j\omega) = -\angle P_{rat}(j\omega)$.*

(ii). If $\phi(j\omega) = 90°$ then we say that the plant and controller are completely misaligned. *Complete misalignment implies that $C(j\omega)P(j\omega) = 0$.*

□

We say loosely that the controller and plant are *well aligned* if $\phi(j\omega) \approx 0°$, and *poorly aligned* if $\phi(j\omega) \approx 90°$.

If the plant and controller are significantly misaligned, then feedback properties at the plant output will be much worse than those at the plant input. Furthermore, closed loop interactions may be very large.

Proposition 3.3 *The output complementary sensitivity function satisfies*

$$\|T_O(j\omega)\| = \frac{|T_I(j\omega)|}{\cos\phi(j\omega)}. \tag{3.2}$$

Furthermore,

$$|T_{O12}(j\omega)| + |T_{O21}(j\omega)| \geq |T_I(j\omega)| \tan\phi(j\omega). \tag{3.3}$$

□

It follows from Proposition 3.3 that poor alignment implies that feedback properties at the plant output are worse than those at the plant input. However, because $\cos\phi$ is relatively insensitive to small deviations of ϕ from zero, it follows that modest amounts of misalignment can be tolerated. Furthermore, if $|T_I(j\omega)|$ is small at a given frequency, then poor alignment at that frequency may not imply poor feedback properties. In particular, it is not necessary to require very good alignment outside the closed loop bandwidth.

---

[1] If $P(s)$ or $C(s)$ has a pole at $s = j\omega$, then one may define direction using the numerator of a coprime fraction.

4 Perfect Plant/Controller Alignment

Although it is not critical that the controller be perfectly aligned with the plant at all frequencies, it is interesting to study perfectly aligned controllers as a limiting case. Hence, suppose that $\phi(j\omega) = 0, \forall\omega$. We now show that if the direction of the plant varies with frequency, then perfect alignment implies that the input open loop transfer function, $L_I(s)$, will have at least one NMP zero.

Factor the plant as

$$P(s) = \begin{bmatrix} n_{p1}(s) \\ n_{p2}(s) \end{bmatrix} K_P(s), \tag{4.1}$$

where $n_{p1}(s)$, and $n_{p2}(s)$ are coprime polynomials, and $K_P(s)$ is a rational transfer function. Note that the direction of the plant is *constant* if and only if both polynomials $n_{p1}(s)$ and $n_{p2}(s)$ are constants. To proceed, factor the controller as

$$C(s) = K_C(s) \left[n_{c1}(s) \quad n_{c2}(s) \right], \tag{4.2}$$

where $n_{c1}(s)$, and $n_{c2}(s)$ are coprime polynomials, and $K_C(s)$ is a rational transfer function.

Proposition 4.1 *The plant and controller are perfectly aligned at all frequencies if and only if the controller can be expressed as* (4.2)*, where* $n_{c1}(s) = n_{p1}(-s)$, $n_{c2}(s) = n_{p2}(-s)$. ☐

Denote the transfer functions associated with a perfectly aligned controller by, e.g., $L_I^*(s), T_O^*(s)$, etc.

Corollary 4.2 (NMP Zeros of $L_I^*(s)$) *Suppose that $P(s)$ and $C(s)$ are perfectly aligned at all frequencies. Then, unless the direction of the plant is constant with frequency, $L_I^*(s)$ must* necessarily *possess at least one NMP zero.*

Proof: Proposition 4.1 implies that $L_I^*(s) = K_C(s) K_P(s) \Delta^*(s)$, where

$$\Delta^*(s) \triangleq n_{p1}(-s) n_{p1}(s) + n_{p2}(-s) n_{p2}(s). \tag{4.3}$$

Unless both $n_{p1}(s)$ and $n_{p2}(s)$ are constant, $\Delta^*(s)$ will have zeros distributed symmetrically about the $j\omega$-axis. The assumption that $n_{p1}(s)$ and $n_{p2}(s)$ are coprime implies that $\Delta^*(s)$ has no zeros on the $j\omega$-axis itself. Hence $\Delta^*(s)$ and $L_I^*(s)$ must have at least one zero lying in the ORHP. ☐

The significance of Corollary 4.2 is that a NMP zero of $L_I(s)$ will impose design limitations upon $S_I(s)$. For example, as discussed in [1], such a zero will limit the achievable closed loop bandwidth.

5 Plant Direction vs Frequency

We showed in Section 4 that requiring perfect alignment at all frequencies introduces unwanted NMP zeros unless the direction of the plant is constant with frequency. By the remarks following Proposition 3.3, it follows that perfect alignment at all frequencies is not necessary to obtain a good feedback design. Hence, let us relax the requirement of perfect alignment, and require only *approximate* alignment over a *finite* frequency range. It seems reasonable to conjecture that design difficulties will arise if the plant direction varies significantly over the frequency range in which requirements are imposed. To expedite our study of this conjecture, in this section we introduce a coordinate system that describes how the plant direction varies with frequency.

Definition 5.1 (DC Coordinates) *Choose unit vectors* $\vec{P}(0)$ *and* $\vec{P}^{\perp}(0)$ *to span* $\mathcal{R}(P(0))$ *and* $\mathcal{N}_{left}(P(0))$, *respectively, and write*

$$P(s) = \vec{P}(0)\tilde{p}_1(s) + \vec{P}^{\perp}(0)^T\tilde{p}_2(s). \tag{5.1}$$

Define an orthogonal matrix

$$M \triangleq \left[\vec{P}(0) \quad \vec{P}^{\perp}(0)^T\right]. \tag{5.2}$$

The transfer functions of the plant, controller, and output complementary sensitivity function in DC coordinates are related to those in standard coordinates by $P(s) = M\tilde{P}(s)$, $C(s) = \tilde{C}(s)M^{-1}$, *and* $T_O(s) = M\tilde{T}_O M^{-1}$. ☐

It is easy to show that transformation to DC coordinates preserves the transfer functions at the plant input (e.g., $L_I(s) = \tilde{L}_I(s)$), the norms of the transfer functions at the plant output (e.g., $\|T_O(s)\| = \|\tilde{T}_O(s)\|$), and the alignment angle, $\phi(j\omega) = \tilde{\phi}(j\omega)$.

By construction, the plant ratio in DC coordinates, given by $\tilde{P}_{rat}(s) = \tilde{p}_2(s)/\tilde{p}_1(s)$, satisfies

$$\tilde{P}_{rat}(0) = 0. \tag{5.3}$$

Moreover, if the plant direction is *constant* with frequency, then $\tilde{P}_{rat}(j\omega)$ is *identically* zero. We shall use the value of $|\tilde{P}_{rat}(j\omega)|$ at a given frequency as a measure of the extent to which the direction of the plant at that frequency has varied from the plant direction at DC.

6 An Analytic Constraint Imposed by Varying Plant Direction

We now derive a design limitation that must be satisfied by the output complementary sensitivity function of any SITO feedback system whose plant direction changes with frequency. To do so, we shall work in the transformed "DC directional" coordinates defined in Section 5.

By analogy with (4.1), factor the plant in transformed coordinates as

$$\tilde{P}(s) = \begin{bmatrix} \tilde{n}_{p1}(s) \\ \tilde{n}_{p2}(s) \end{bmatrix} \tilde{K}_P(s), \tag{6.1}$$

where $\tilde{n}_{p1}(s)$ and $\tilde{n}_{p2}(s)$ are coprime polynomials. Denote the

(i). NMP zeros of $\Delta^*(s)$ by $\{\delta_i : i = 1 \ldots N_\delta\}$,

(ii). NMP zeros of $\tilde{n}_{p1}(s)$ by $\{\rho_i : i = 1 \ldots N_\rho\}$,

(iii). OLHP zeros of $\tilde{n}_{p1}(s)$ by $\{\lambda_i : i = 1 \ldots N_\lambda\}$.

We now show that $\|T_O(j\omega)\|$ is constrained by an integral relation; for a proof, see [5].

Proposition 6.1 *Assume that the feedback system is stable. Then*

$$\int_0^\infty \log\left(\frac{\|T_O(j\omega)\|}{|T_I(0)|}\right)\frac{d\omega}{\omega^2} \geq F(P) + \frac{\pi}{2}\lim_{s\to 0}\frac{d}{ds}\frac{T_I(s)}{T_I(0)} - \frac{\pi}{2}\tan\phi(0)\left|\lim_{s\to 0}\frac{d}{ds}\tilde{P}_{rat}(s)\right|, \tag{6.2}$$

where

$$F(P) \triangleq \frac{\pi}{2}\left(\sum_{i=1}^{N_\delta} \frac{1}{\delta_i} + \sum_{i=1}^{N_\rho} \frac{1}{\rho_i} - \sum_{i=1}^{N_\lambda} \frac{1}{\lambda_i}\right). \tag{6.3}$$

Suppose, in addition, that $\tilde{P}_{rat}(s)$ has at least two zeros at $s = 0$ and that $L_I(s)$ has at least two integrators. Then

$$\int_0^\infty \log\left(\frac{\|T_O(j\omega)\|}{|T_I(0)|}\right)\frac{d\omega}{\omega^2} \geq F(P). \tag{6.4}$$

□

It is clear from (3.2) that $\|T_O(0)\| \geq |T_I(0)|$, and thus it is trivially true that the ratio $\|T_O(j\omega)\|/|T_I(0)|$ must exceed one at some frequency. The lower bound (6.2) suggests that there may be frequencies for which $\|T_O(j\omega)\| \gg |T_I(0)|$. Since good design practice usually requires that $|T_I(0)| \approx 1$, it follows that feedback properties at the plant output will be poor. To obtain an explicit lower bound on the peak in $\|T_O(j\omega)\|$ requires an additional hypothesis, and will be addressed in Corollary 6.4 below. We first describe the size of each of the three terms on the right hand side of (6.2).

Note first that $F(P)$ depends only upon the plant, and has no analog for SISO systems. Additional properties follow.

Lemma 6.2 *Let $F(P)$ be defined by* (6.3). *Then $F(P) \geq 0$, and $F(P) = 0$ if and only if the plant direction is constant with frequency.*

Assume next that

$$|\tilde{P}_{rat}(j\omega)|^2 \leq \frac{\epsilon\omega^2}{\omega^2 + \omega_0^2}, \quad \forall\omega, \quad \epsilon < 1/2. \tag{6.5}$$

Then

$$F(P) \leq \frac{3\pi\epsilon}{4\omega_0}. \tag{6.6}$$

Proof: The first claim follows immediately from the fact that N_δ, N_ρ, and N_λ are all trivially equal to zero if the plant direction is constant. The second claim follows from the identity

$$\frac{1}{2}\int_0^\infty \log\left(1 + |\tilde{P}_{rat}(j\omega)|^2\right)\frac{d\omega}{\omega^2} = \frac{\pi}{2}\left(\sum_{i=1}^{N_\delta} \frac{1}{\delta_i} + \sum_{i=1}^{N_\rho} \frac{1}{\rho_i} - \sum_{i=1}^{N_\lambda} \frac{1}{\lambda_i}\right). \tag{6.7}$$

The proof of (6.6) is similar to that of Corollary 3.1.6 in [3]. □

It follows from Lemma 6.2 that if the plant direction varies with frequency, then the lower bound for the special case (6.4) will be strictly greater than one. However, if the frequency variation is small, in the sense of (6.5), then $F(P)$ will also be small.

The second term on the right hand side of (6.2) is present in SISO systems, for which it plays a role analogous to the reciprocal of the velocity constant in a Type-1 feedback system [3]. In the present case, note that [4]

$$\lim_{s\to 0}\frac{d}{ds}T_I(s) = \sum_{i=1}^{N_p} \frac{1}{p_i} - \sum_{i=1}^{N_z} \frac{1}{z_i}, \tag{6.8}$$

where $\{z_i : i = 1, \ldots, N_z\}$ and $\{p_i : i = 1, \ldots, N_p\}$ are the zeros and poles of $T_I(s)$. It follows that this term can be large and negative only if $T_I(s)$ contains one or more NMP zeros, has high dynamical order, and/or contains slow stable poles that do not have nearby zeros. These properties all tend to limit the bandwidth of the closed loop system.

The third term on the right hand side of (6.2) also has no counterpart for SISO systems. The second factor, $\lim_{s\to 0} d\tilde{P}_{rat}(s)/ds$, is a fixed property of the plant. The first factor, $\tan\phi(0)$, can be adjusted to make this term arbitrarily large, but only at the expense of large misalignment at DC, with attendant poor feedback properties.

We now demonstrate that if the direction of the plant changes with frequency, then the feedback system faces design limitations that are more severe than if the direction of the plant were constant. To do so, we shall construct a set of design specifications that may be satisfied by a stable, minimum phase SISO plant. These specifications may also be satisfied for a stable, minimum phase SITO plant *whose direction is constant with frequency.*

Define $L(s) = \omega_0/s$, and $T(s) = L(s)/(1 + L(s)$, and suppose that $T_I(s)$ and $T_O(s)$ satisfy:

$$\begin{aligned}
\lim_{s\to 0} \frac{d}{ds}\frac{T_I(s)}{T_I(0)} &\geq \lim_{s\to 0} \frac{d}{ds}\frac{T(s)}{T(0)} \\
&= \frac{-1}{\omega_0}, && (6.9)\\
\|T_O(\omega)\| &\leq |T(j\omega)| \\
&= \left|\frac{\omega_0}{j\omega + \omega_0}\right|, \quad \forall\omega \leq \omega_0, && (6.10)\\
\phi(0) &\leq \phi_{max}, && (6.11)\\
\|T_O(j\omega)\| &\leq T_O^{max}, \forall\omega. && (6.12)
\end{aligned}$$

It follows by construction that if a SISO plant is stable and minimum phase, then a (possibly improper) stabilizing controller may be found so that the complementary sensitivity function $T(s)$ will satisfy specifications of the form (6.9)-(6.10) for ω_0 arbitrarily large and with $\|T\|_\infty \leq 1$.

If the direction of a SITO plant is constant (and there are no other sources of design limitations) then our next result shows that the specifications (6.9)-(6.11) can also be achieved for arbitrary ω_0 and with $\|T_O\|_\infty \leq 1$.

Proposition 6.3 *Assume that the plant has the form*

$$P(s) = \begin{bmatrix} n_{P1} \\ n_{P2} \end{bmatrix} K_P(s) \tag{6.13}$$

where n_{P1} and n_{P2} are constants and $K_P(s)$ is stable and minimum phase. Then for any value of ω_0 there exists a (possibly improper) stabilizing controller such that (6.9)-(6.10) *are satisfied, $\|T_O\|_\infty \leq 1$, and $\phi_{max} = 0$.*

Proof: The controller

$$C(s) = \frac{\omega_0}{s}\frac{1}{K_P(s)}\frac{[n_{P1} \;\; n_{P2}]}{\|[n_{P1} \;\; n_{P2}]\|^2} \tag{6.14}$$

yields

$$T_I(s) = \frac{\omega_0}{s + \omega_0}, \tag{6.15}$$

$$\|T_O(\omega)\| = \left|\frac{\omega_0/j\omega}{1 + \omega_0/j\omega}\right|, \;\; \forall\omega \tag{6.16}$$

and the result follows. □

On the other hand, if the direction of a SITO plant varies with frequency, then Proposition 6.1 may prevent specifications of the form (6.9)-(6.11) from being satisfied without forcing large peaks in $\|T_O\|_\infty$.

Corollary 6.4 *Assume that* (6.9)-(6.11) *are satisfied. Then*

$$\begin{aligned}\log\frac{\|T_O\|_\infty}{|T_I(0)|} &\geq \omega_0 F(P) - \omega_0\int_0^{\omega_0}\log\left|\frac{\omega_0/j\omega}{1+\omega_0/j\omega}\right|\frac{d\omega}{\omega^2}\\ &\quad -\frac{\pi}{2}-\frac{\pi}{2}\omega_0\tan\phi_{max}\left|\lim_{s\to 0}\frac{d}{ds}\tilde{P}_{rat}(s)\right| \end{aligned}\tag{6.17}$$

□

Note that many other scalar transfer functions could be used to construct specifications (6.9)-(6.10). We shall illustrate the bound (6.17) in the next section.

7 Examples

We now consider two examples of SITO plants for which the plant direction changes with frequency. In Example 7.1 the plant direction changes significantly with frequency, resulting in a nontrivial bandwidth limitation for the closed loop system. In Example 7.2 the plant direction does not change significantly with frequency and the closed loop system can achieve much larger bandwidth without causing large values of $\|T_O\|_\infty$.

Example 7.1 Consider the plant

$$P(s) = K_P(s)\begin{bmatrix}1\\ s\end{bmatrix}, \tag{7.1}$$

where $K_P(s)$ is stable and minimum phase.

The lower bound (6.17) is plotted in Figure 1 for various values of $\tan\phi_{max}$. Note that the lower bound on $\|T_O\|_\infty/|T_I(0)|$ becomes very large as the value of ω_0 increases. These bounds indicate that as ω_0 increases, requiring $|T_I(0)| \approx 1$ will force $\|T_O\|_\infty$ to become very large. Therefore the achievable value of ω_0 is bounded above due to the change in the direction of the plant with frequency. □

Example 7.2 Consider a plant

$$\begin{aligned}P(s) &= \begin{bmatrix}(1+\epsilon)s+1\\(1-\epsilon)s+1\end{bmatrix}K_P(s),\\ &= (s+1)\begin{bmatrix}1\\1\end{bmatrix}K_P(s) + \epsilon s\begin{bmatrix}1\\-1\end{bmatrix}K_P(s).\end{aligned}\tag{7.2}$$

It is easy to verify that $\Delta^*(s)$ has a NMP zero at $s = 1/\sqrt{1+\epsilon^2}$. This zero will limit the closed loop bandwidth if the controller is perfectly aligned with the plant. In DC coordinates, $\tilde{n}_{p1}(s)$ has an OLHP zero at $s=-1$. It is easy to verify that the direction of the plant does not change very much with frequency, and one suspects that the attendant design limitation is not severe if we introduce a modest amount of misalignment.

Let $\epsilon = 0.01$. In this case $F(P) \approx \frac{\pi}{2}5\times 10^{-5}$, and is thus not very large. The lower bound (6.17), plotted for $\phi(0)=0$, does not imply a peak in $\|T_O\|_\infty/|T_I(0)|$ for relatively large values of ω_0, and suggests that it will not be very difficult to satisfy the specifications (6.9)-(6.12) with $\omega_0 \approx 100 rad/sec$.

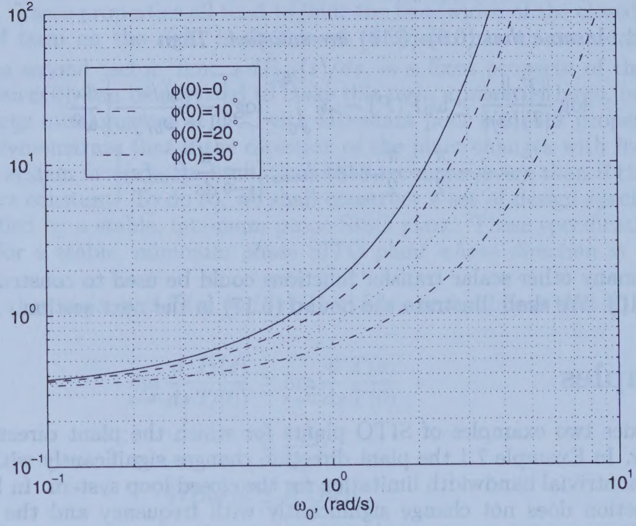


Figure 1: Lower bound on $\|T_O\|_\infty/|T_I(0)|$, Example 1

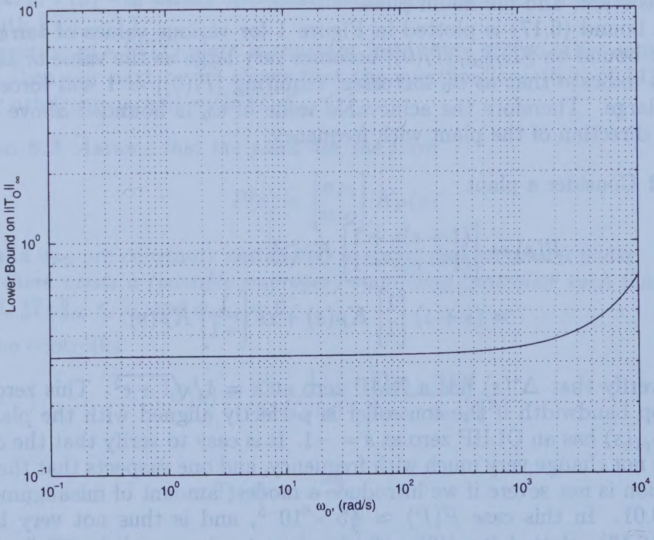


Figure 2: Lower bound on $\|T_O\|_\infty/|T_I(0)|$, Example 2

Consider a controller that is aligned with the DC gain of the plant

$$C(s) = \frac{1}{(s+1)K_P(s)}\frac{\omega_b}{s}\frac{k}{\tau s+1}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}. \tag{7.3}$$

Choosing parameter values $\omega_b = 100, \tau = 0.0001, k = 0.7$ yields a closed loop system that is stable. The -3dB frequency for both the input and output complementary sensitivity functions is approximately $100 rad/sec$ and $\|T_O\|_\infty \leq 1$. This shows that there is indeed no obstacle to obtaining high closed loop bandwidth for this plant whose direction varies only slightly with frequency. □

In each case, requiring perfect alignment at all frequencies yields a bandwidth limitation due to a NMP zero of $L_I^*(s)$ at approximately $s = 1$. For the first example, the change in plant direction does indeed impose constraints on the achievable performance of the closed loop system. In the second case, the constraints are much less severe. Therefore using perfectly aligned controllers when the plant does not vary significantly with frequency can result in performance greatly inferior to controllers that are only slightly misaligned with the plant.

8 Summary and Conclusions

In this paper we have discussed the importance of alignment in the control of SITO systems. We have shown that requiring perfect alignment for systems which vary greatly with frequency implies a bandwidth constraint on the closed loop system. We have also shown that requiring perfect alignment at all frequencies may be undesirable in the control of SITO systems if the plant direction does not change significantly with frequency. This design limitation ihas no counterpart in SISO systems.

References

[1] J. S. Freudenberg and D. P. Looze. *Frequency Domain Properties of Scalar and Multivariable Feedback Systems*, volume 104 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, 1988.

[2] J. S. Freudenberg and R. H. Middleton. Design rules for multivariable feedback systems. In *Proceedings of 35th Conference on Decision and Control*, Kobe, Japan, December 1996.

[3] M. M. Seron, J. H. Braslavsky, and G. C. Goodwin. *Fundamental Limitations in Filtering and Control.* Springer, 1997.

[4] John G. Truxal. *Automatic Feedback Control System Synthesis.* Mc-Graw Hill, 1955.

[5] A. R. Woodyatt, J. S. Freudenberg, and R. H. Middleton. A design tradeoff for single input two output feedback systems. Technical Report EE9703, University of Newcastle, 1997.

Topological Concepts for Robustness Margin Problems

Nainn-Ping Ke
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, CA 90089-2563
email: nke@eudoxus.usc.edu

Abstract

The objective of this paper is applying some topological concepts to enhance the Horowitz template approach for analyzing the robustness properties of linear feedback system. We choose two well-known real robustness margin problems as examples to demonstrate the usage of these concepts. It is shown that these concepts are useful to understand the discontinuity phenomenon of real robustness margin problems.

1 Introduction

This paper deals with a stability robustness problem. Stability robustness, which every feedback control system should have in order to remain stable for all conditions, is the most important issue of a feedback controller design. In this paper, we consider the robustness margin problem of plants with real/complex parametric uncertainty. This problem can be formulated in terms of a canonical diagonal perturbation [5, 20, 21]. An algorithm for exact calculation of the multiloop stability margin for real uncertainty has been proposed by de Gaston and Safonov [5]. Structured singular value, which is another way to formulate this problem, is proposed by Doyle [4] and he used upper and lower bound to approximate this structured singular value. It is well-known that the real robustness margin can be discontinuous in the problem data, while the complex robustness margin is always continuous [2, 8, 10, 11, 18]. The discontinuity of the real robustness margin casts some doubt on the usefulness of the real robustness margin. In the mean time, the complex robustness margin is criticized for being too conservative. This is the main motivation to study robustness margin problem.

It has been shown that singularity theory is useful to understand this discontinuity phenomenon [9, 10, 12, 13, 16, 22]. Instead of using the upper bound and lower bound approach, this paper uses the Horowitz template approach to capture the geometric/differential topological properties of the stability margin. We apply differential topological concepts, such as stable property, transversality and stable mapping to two well-known real robustness margin problems to explain why the lack of continuity of structured stability margin relative to problem data could happen. The transversality discussed in this paper have been used in analyzing the robustness properties of linear feedback system [9, 12]. Based on these topological concepts, stratified Morse theory and open covering concept, a new algorithm, boundary approach, of symbolic computation for exact stability margin is proposed in [15].

2 Problem Formulation

We need to formulate the robust margin problem within a framework amenable to mathematical analysis.

Definition 1 (Nyquist Map and Nyquist template) *The Nyquist map is defined as [9]*

$$\begin{aligned} f:\ \mathbb{D}\times\Omega &\longrightarrow\ \mathbb{C} \\ (q,j\omega) &\longmapsto\ det(I+L(q,j\omega)) \end{aligned}$$

where the domain $\mathbb{D}\times\Omega$ *is the space of uncertain parameters* $\mathbb{D}$ *(e.g., the Kharitonov cube) topologically multiplied by the frequency spectrum* Ω*.* $L(q,s)\in\mathbb{C}^{n\times n}$ *is the stable loop transfer matrix.* q *is the vector of parametric uncertainty. The Nyquist template is* $N=f(\mathbb{D}\times\Omega)$*. The Horowitz template is* $N_\omega=f(\mathbb{D},\omega)$*.*

Definition 2 (Robustness Margin) *The maximum uncertainty bound,* r^c_{max}*, such that the system is stable for all* $q\in\mathbb{C}^n$ *with* $\|q\|_\infty\le r^c_{max}$ *is called the real/complex robustness margin of a control system. In the above*

$$\|q\|_\infty \stackrel{\text{def}}{=} \max_i\{\mid q_i\mid\}$$

where $\mid q_i\mid$ *denotes, in the real case, the absolute value of* q_i *and, in the complex case, the modulus of* q_i*.*

Now, the robustness margin problem with real parametric uncertainties becomes that of finding the smallest $\|\ q\ \|_\infty$ such that the template hits (0+j0).

$$r_{max}(\omega)=\inf\{\|\ q\ \|_\infty\colon det(I+L(q,\omega))=0\}$$

$$r_{max}=\inf_\omega\{r_{max}(\omega)\}$$

Instead of using the upper bound and lower bound approach [5], this paper uses the critical values plots to capture the geometric/differential topological properties of the stability margin. Structured Singular value, μ, introduced in [4], is an equivalent way to formulate this problem where $\mu=1/r_{max}$.
In [9, 18], it was shown that r_{max} with complex parametric uncertainty is continuous relative to rounding errors. On the contrary, r_{max} is discontinuous with real uncertainty [2]. As shown in [9], this discontinuity phenomenon is due to $0+j0\in\mathbb{C}$ being a critical value of the Nyquist map. In many cases, real parametric uncertainty is a more realistic formulation.

In this paper, we consider the Nyquist map as a general map, not restricted to be a multilinear map [5, 9].

3 Morse Theory and Some Topological Concepts

3.1 Morse Theory

The Morse theory is an important tool to understand the boundary of the Horowitz template. The boundary of the template is crucial in the analysis of the robust stability problem [9]. The boundary of the template is first to intercept 0+j0 when $\|q\|_\infty$ is increased.

Originally the Morse theory was developed for compact, differentiable manifolds [17]. Unfortunately, in many structured stability problems, the set of uncertainties $\mathbb{D}$ is more complicated than a mere compact differentiable manifold. For example, the (closed) Kharitonov cube is not a manifold [9]. The basic theme of the Morse theory is to establish the connections between the critical points of a real-valued function and the topology of its domain. Some important terminology is as follows [9, 14, 17]:

Definition 3 (Critical Point) *A critical point q^* of a smooth map f_ω defined over a smooth manifold $\mathbb{D}$ is a real point where the induced map*

$$df_\omega : T_{q^*}\mathbb{D} \longrightarrow \mathbb{R}^2$$

is not surjective, that is, $dim_{\mathbb{R}}\ df_\omega(T_{q^*}\mathbb{D}) < 2$. A point in the domain which is not a critical point is a regular point. A critical value is the value of the map, $f(q^*)$, at a critical point. A point in the range which is not a critical value is called regular value.

Using local Riemann coordinates yields a more intuitive definition.

Definition 4 (Critical Point) *A critical point q^* of a smooth map f_ω defined over a smooth manifold is a point where the rank, over the ground field $\mathbb{R}$, of the Jacobian representation of df_ω,*

$$J^s = \begin{pmatrix} \frac{\partial \Re f_\omega}{\partial q_1} & \frac{\partial \Re f_\omega}{\partial q_2} & \cdots & \frac{\partial \Re f_\omega}{\partial q_n} \\ \frac{\partial \Im f_\omega}{\partial q_1} & \frac{\partial \Im f_\omega}{\partial q_2} & \cdots & \frac{\partial \Im f_\omega}{\partial q_n} \end{pmatrix} \tag{1}$$

is < 2.

Theorem 1 (Sard's Theorem) *If $f_\omega : X \longrightarrow Y$ is a smooth map between smooth manifolds, then almost every point in Y is a regular value of f_ω [7]. In other words, if $f_\omega : X \longrightarrow Y$ is a smooth map, the set of critical values has Lebesgue measure zero.*

The rank of the Jacobian may drop to either 1 or 0.

Definition 5 (Boundary) *As a subset N of $\mathbb{R}^2$, ∂N, the boundary of N , is defined as $\partial N = \overline{N} - Int(N)$. The closure of N, $\overline{N}$, is the intersection of all closed sets of $\mathbb{R}^2$ covering N, while the interior of N, $Int(N)$, is the union of all open sets contained in N.*

3.2 Topological Concepts

The robustness margin need not to be continuous relative to some problem data is a well-known result in [2, 14]. This phenomenon is hard to explain. Also, to solve the example in [1] has some difficulty in regarding the regularity. In this paper, we want to use the following differential topological tools to analyze these problem and answer why discontinuity happens in real robustness problem - from geometric point of view. In the mean time, we like to compare our results with algebraic results.

The preimage of a regular value of the smooth map $f : X \longrightarrow Y$ is a submanifold of X. This simple fact has led mathematician to a generalization of the notion of regularity-namely, transversality [6, 7].

Definition 6 (Stable Property) *A property is stable provided that whenever* $f_0 : X \longrightarrow Y$ *possesses the property and* $f_t : X \longrightarrow Y$ *is a homotopy of* f_0*, then for some* $\epsilon > 0$*, each* f_t *with* $t < \epsilon$ *also possesses the property.*

From def 6, we have the following observations:

- The property that a curve pass through the origin in a plane, i.e., $\mathbb{R}^2$ is not stable, since any small deformation can distort this curve to avoid the origin [7]. This is shown in Figure 1.
- A continuous injective map from an open square $(-1, 1)^2$ into $\mathbb{C}$. The property that the image of this map pass through the origin is stable.
- Our robustness margin problem will have a simple question need to be answer "Is the Horowitz template intersect the origin stable?".

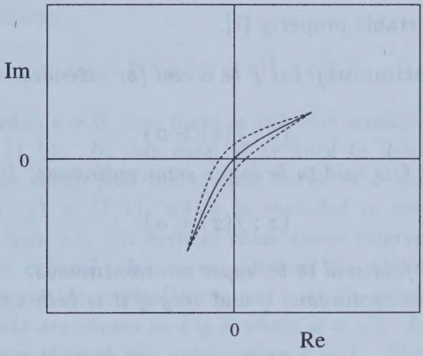


Figure 1: The property of a curve pass through the origin of a plane is not stable.

Definition 7 (Stable Map) *A smooth map* $f : X \longrightarrow Y$ *between smooth manifolds* X, Y *is said to be stable whenever there exists a neighborhood* O_f *of* f *in* $C^\infty(X, Y)$ *equipped with the Whitney topology such that* $\forall f' \in O_f$ *there exist diffeomorphisms* h *and* g *such that the following diagram commutes.*

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "f"] \arrow[d, "h"] & \mathcal{Y} \arrow[d, "g"] \\
\mathcal{X} \arrow[r, "f'"] & \mathcal{Y}
\end{tikzcd}$$

To answer the previous question, we need the following transversality condition.

Definition 8 (Transversality) *Let* $f : X \longrightarrow Y$ *be a smooth map between smooth manifolds. Let* $Z \subseteq Y$ *be a submanifold. Then* f *is said to intersect* Z *transversally iff*

$$T_{f(x)}Y = (d_x f)(T_x X) + T_{f(x)}Z, \text{ for } \forall x \in f^{-1}(Z) \tag{2}$$

where, typically, $T_x X$ *denotes the tangent space to* X *at* x*.*

From above definitions, we have the following observations:

- If $f(X) \cap Z = \emptyset$, then f transverse Z.
- A injective smooth map $f : \mathbb{R} \longrightarrow \mathbb{C}$, i.e. a curve, never intersect the origin in $\mathbb{C}$ transversally except when they do not intersect at all.
- In robustness control problem, we concern about the robustness margin which is related to the preimage of the origin. In this case, Z is the origin and its tangent space $T_{f(x)}Z = 0$. Therefore, the transversal condition, i.e. Eq. 2, reduce to be

  $$T_{f(x)}Y = (d_x f)(T_x X), \forall x \in f^{-1}(0 + j0)$$

  Therefore, the Nyquist map f intersects the origin transversally iff the induced map $d_x f : T_x \longrightarrow T_{f(x)}Y$ is surjective, i.e. f is a submersion at x. Therefore, the necessary condition of transversality is that $\forall x \in f^{-1}(0 + j0)$ are regular points. Furthermore, the Nyquist map f does not intersect the origin transversally if the critical value curve hit the origin. Moreover, the discontinuity of robustness margin may happens when the transversal condition be violated.
- Transversality is a stable property [7].

Definition 9 (Semicontinuous) *Let f be a real (or extended-real) function on a topological space. If*

$$\{x : f(x) > \alpha\}$$

is open for every real α, f is said to be lower semicontinuous. If

$$\{x : f(x) < \alpha\}$$

is open for every real α, f is said to be upper semicontinuous.
A real function is obvious continuous if and only if it is both upper and lower semicontinuous [19].

Lemma 2 *μ is always an upper semicontinuous function [18].*

4 Real versus Complex Robustness

When we model the uncertainties of a given system by real instead of complex parameters the resulting robustness margin is so different. This is related to the property of the image. The concepts of open and closed set play a key role in the subject of topology. These concepts also play a important role in robustness margin problem. In [11], it point out due to open mapping property of several complex variables if we model the uncertainty by complex the resulting robustness is continuous. Here we summarize the different property of real and complex uncertainty as follows tables [11, 12]:

Uncertainty Type	Resulting Robustness Margin	Image(Domain is a Open Set)
Real	Could Be Discontinuous	Could be not open
Complex	Is Always Continuous	Always Open

Uncertainty Type	Number of Constraints	Type of Critical Points
Real	n-1	Generically Is a Critical Curve
Complex	n	Finite Critical Points

Uncertainty Type	Resulting Image
Real	Not Always Surjective
Complex	Surjective

5 Examples

The following examples are typical problems of real robustness margin problems. The robustness of these problems is discontinuous at some frequency relative to problem data. We will apply the topological concepts to these problems and do simulation to show the usefulness of the topological concepts. Due to "0/0 Ambiguity" problem [10], the open-loop transfer function of Example 2 in [2] is modified.

Example 1 *Consider the following example [1]:*

$$P(q,s) = s^3 + (q_1 + q_2 + 1)s^2 + (q_1 + q_2 + 3)s + (1 + \epsilon + 6q_1 + 6q_2 + 2q_1 q_2). \tag{3}$$

The unstable region in the uncertainty plane can be calculated from Routh-Hurwitz array which is described as follows:

$$(q_1 - 1)^2 + (q_2 - 1)^2 - \epsilon \leq 0.$$

The major problem is when $\epsilon = 0$*, then there is only one unstable point in the uncertainty plane, i.e.* $(q_1, q_2) = (1,1)$*. In this case, it is hard to detect this unstability. Use the definition 4, we can easily find the critical curve is a line, i.e.,* $q_1 = q_2$ *at* $\omega = \sqrt{5}$*. The image of* $(q_1, q_2) = (1,1)$*, which is included in critical value curve, is the origin of the complex plane, i.e. the critical value curve intersect the origin. As earlier discussion indicated, the critical value curve intersect the origin is not a stable property any disturbance could cause the critical curve not pass the origin.*

The simulation results are shown in Fig 2 where $\omega = \sqrt{5}$*. In Figure 2(a) the critical value curve does not pass through the origin when* $\epsilon = 1$*. Therefore, the critical value curve and the origin intersect transversally in Fig 2(a). In Figure 2(b) the critical value curve passes through the origin when* $\epsilon = 0$*. Therefore, the critical value curve and the origin do not intersect transversally in Figure 2. When* $\epsilon < 0$*, the image will not intersect the origin and the system is stable in this case. These results also can be explained by the concept of stable property, i.e., the property that the critical value curve pass through the origin is not stable.*

Example 2 *Consider a unity feedback system with open-loop transfer function [2, 3, 11, 13]:*

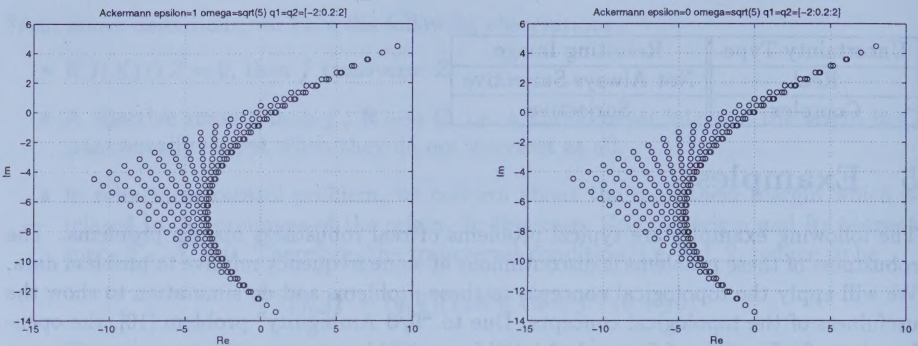
$$L_n(q,s) = \frac{1}{D(q,s)} \tag{4}$$

where

$$\begin{aligned} D(q,s) &= s^4 + (20 - 20q_2)s^3 + (44 + 2a + 10q_1 - 40q_2)s^2 + (20 + 8a + 20aq_1 - 20q_2)s \\ &\quad + a^2 + k(4a + 10aq_1) - 1 \end{aligned}$$

$$k = (a + \gamma)$$

$$a = 3 + 2\sqrt{2}$$



(a)Example 1 - stable case. (b)Example 1 - unstable case.$k = a$

Figure 2: Stable and unstable

The closed-loop characteristic polynomial is

$$\begin{aligned} g(q,s) &= s^4 + (20 - 20q_2)s^3 + (44 + 2a + 10q_1 - 40q_2)s^2 \\ &\quad + (20 + 8a + 20aq_1 - 20q_2)s + a^2 + k(4a + 10aq_1) \end{aligned} \tag{5}$$

The robustness margin of the counterexample can be computed by the frequency sweep method or the Routh-Hurwitz array [2, 13]. Use the def 4, we can easily find out there is no critical point when $\gamma \neq 0$ *at* $\omega = \sqrt{a}$ *and whole uncertainty plane is critical when* $\gamma = 0$ *at* $\omega = \sqrt{a}$. *Therefore, as earlier discussion indicated, the robustness margin is sensitive to problem data when* $\gamma = 0$ *at* $\omega = \sqrt{a}$. *Moreover, the robustness margin* $r_{max} = \inf_\omega \{r_{max}(j\omega)\}$ *is discontinuous at the data point k=a, i.e.*

1. *When* $k = a + \gamma$ *,where* $\gamma \neq 0$,

$$\lim_{\gamma \to 0} r_{max} \cong 0.417.$$

2. *When k=a,*

$$r_{max} \cong 0.234.$$

For $\omega = \sqrt{a}$, *the* μ *is shown in Figure 3 in which it is a upper semicontinuous function. It is surprising to see the discontinuity at* $\gamma = 0$.

This discontinuity of robustness margin can be explained by the concept of the stable property. In Figure 4, assume the image is Horowitz template of an interval [-1,1]. The Figure 4(a) has shown that there is only one point of the image mapped to the origin. If we distort this curve we could get Figure 4(b) which has two points mapped to the origin. If we perturbed this curve again we could get Figure 4(c) which has only one point mapped to the origin. If we assume there is a small perturbation to obtain the Figure 4(b) from the Figure 4(a). Therefore, the μ *of the Figure 4(b) could increase a lot. If we also assume the Figure 4(c) is obtained from the Figure 4(b) with a small perturbation. The* μ *of the Figure 4(c) could decrease a lot from the* μ *of the Figure 4(b). The simulation results of the Example 2 are shown in Figure 5(a) where* $\gamma = -1$ *and* $\omega = \sqrt{a}$ *and Figure 5(b) where* $\gamma = 0$ *and* $\omega = \sqrt{a}$. *There is no critical points when* $\gamma \neq 0$. *The image of Nyquist map of Example 5 intersect the origin transversally when* $\gamma \neq 0$ *at* $\omega = \sqrt{a}$ *which is*

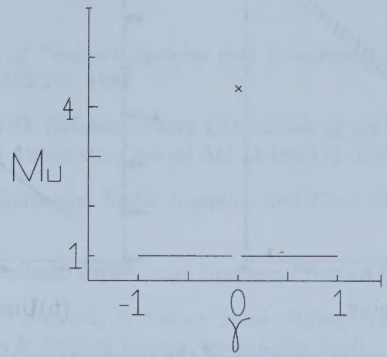
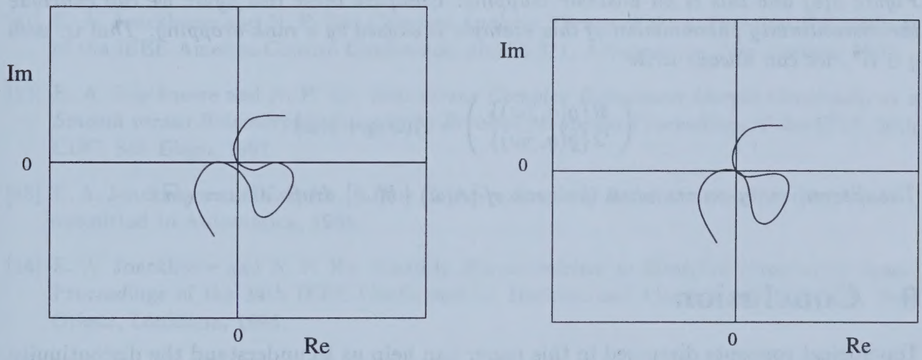
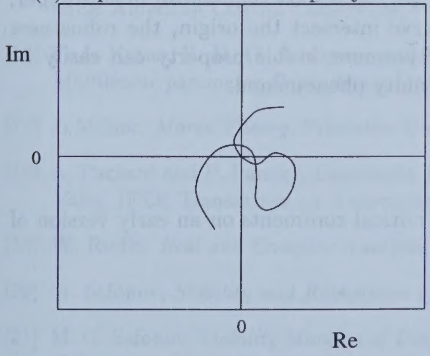


Figure 3: Mu vs. Gamma

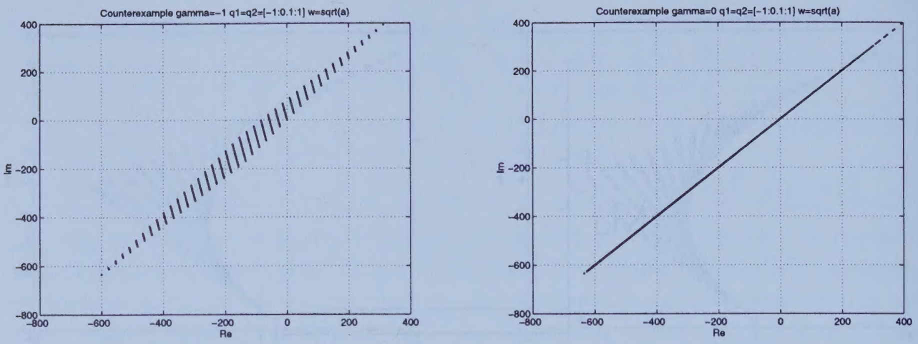


(a)Map with Unstable Property - 1 (b)Map with Unstable Property - 2



(c)Map with Unstable Property - 3

Figure 4: Stable and unstable



(a)Stable mapping (b)Unstable mapping

Figure 5: Horowitz template of the counterexample at $\omega = \sqrt{a}$

shown in Figure 5(a) and this is a stable mapping. When $\gamma = 0$ *at* $\omega = \sqrt{a}$ *any point in the uncertainty palne is critical points. The image of the Nyquist map of Example 2 do not intersect the origin transversally when* $\gamma = 0$ *and* $\omega = \sqrt{a}$ *which is shown in Figure 5(b) and this is an unstable mapping. Compare these two figure we can conclude the discontinuity phenomenon of this example is caused by a rank-dropping. That is, with* $q \in R^n$*, we can always write*

$$\begin{pmatrix} \Re\{g(q,j\omega)\} \\ \Im\{g(q,j\omega)\} \end{pmatrix} = A(\omega)q + b(\omega)$$

The discontinuity occurs when the rank of $[A(\omega) \mid b(\omega)]$ *drops at* $\omega = \sqrt{a}$.

6 Conclusion

Topological concepts discussed in this paper can help us to understand the discontinuity phenomenon of the robustness margin problem. As the results that shown in this paper, it is easy to understand when critical value curve intersect the origin, the robustness margin are sensitive to the problem data. Furthermore, stable property can easily be understood and be used to explain this discontinuity phenomenon.

Acknowledgement

Many thanks to Prof. E. A. Jonckheere for his critical comments on an early version of this paper.

References

[1] J. Ackermann, A. Bartlett, D. Kaesbauer, W. Sienel and R. Steinhauser, *Robust Control-Systems with Uncertain Physical Parameters*, Springer-Verlag, 1993.

[2] B. Barmish, *New Tools for Robustness of linear System*, Macmillan, 1994.

[3] S. Bhattacharyya, H. Chapellat and L. Keel, *Robust Control-The Parametric Approach*, Prentice Hall, 1995.

[4] J. C. Doyle, *Analysis of Feedback Systems with Structured Uncertainty*, IEE Proc., vol. 129. pt. D. no. 6. pp. 242-250, 1982.

[5] R. de Gaston and M. G. Safonov, *Exact Calculation of the Multiloop Stability Margin*, IEEE Transactions on Automatic Control AC-33:156-171, 1988.

[6] M. Golubisky and V. Guillemin, *Stable Mappings and Their Singularities*, Springer-Verlag, 1973.

[7] V. Guillemin and A. Pollack, *Differential Topology*, Prentice-Hall, 1974.

[8] D. Hinrichsen and A. Pritchard, *A Not on Some Differences between Real and Complex Stability Radii*, System & Control Lettesr, pp.401-408, 1990.

[9] E. A. Jonckheere, *Algebraic and Differential Topology of Robust Stability*, Oxford University Press, 1997.

[10] E. A. Jonckheere and N. P. Ke, *Topological Theory of* 0/0 *Ambiguities in Robust Control*, to appear Proceedings of the IEEE 36th CDC, San Diego, California, 1997.

[11] E. A. Jonckheere and N. P. Ke, *Complex-Analytic Theory of the μ-Function*, Proceedings of the IEEE America Control Conference, pp.366-371, Albuquerque, New Mexico, 1997.

[12] E. A. Jonckheere and N. P. Ke, *Real versus Complex Robustness Margin Continuity as a Smooth versus Holomorphic Singularity Problem*, to appear Proceedings of the IEEE 36th CDC, San Diego, 1997.

[13] E. A. Jonckheere and N. P. Ke, *A New Counterexample of Real Robustness Margin Problem*, submitted to Automatica, 1995.

[14] E. A. Jonckheere and N. P. Ke, *Stability Margin relative to Stratified Uncertainty Space*, Proceedings of the 34th IEEE Conference on Decision and Control, pp.1322-1323, New Orlean, Louisiana, 1995.

[15] N. P. Ke, *An Algorithm of Symbolic Computation for Stability Margin*, submitted to the IEEE American Control Conference, Philadelphia, Pennsylvania, 1998.

[16] F. J. Kraus, B. D. O. Anderson and M. Mansour, *Robust Stability of Polynomials with Multilinear parameter Dependence*, Int. J. Control, pp.1745-1762, 1989.

[17] J. Milnor, *Morse Theory*, Princeton University Press, 1963.

[18] A. Packard and P. Pandey, *Continuity Properties of the Real/Complex Structured Singular Value*, IEEE Transaction on Automatic Control AC-38: 415-428, 1993.

[19] W. Rudin, *Real and Complex Analysis*, New York, McGraw-Hill, 1966.

[20] M. Safonov, *Stability and Robustness of Multivariable Feedback Systems*, MIT, 1980.

[21] M. G. Safonov, *Stability Margins of Diagonally Perturbed Multivariables Feedback Systems*, IEE Pro., vol. 129. Pt. D. no. 6. pp. 242-250,1982.

[22] A. Vicino and A. Tesi, *Regularity Conditions for the Stability Margin Problem with Linear Dependent Perturbations*, SIAM J. Control and Optimizations, 1995.

Solution of the State Feedback Singular H^∞ Control Problem for Linear Time-Varying Systems

F. Amato, M. Mattei and A. Pironti

Dipartimento di Informatica e Sistemistica
Università degli Studi di Napoli "Federico II", via Claudio 21, Napoli 80125
Email: {amato,mattei,pironti}@disna.dis.unina.it

Abstract

In this paper we consider the singular, finite horizon, state feedback H^∞ control problem for linear time-varying systems. Singular problem means that the feedthrough matrix between the control input and the controlled output is not full column rank; in this case we prove that the original H^∞ problem is equivalent to another H^∞ problem related to a *reduced order* system. If this problem is regular it can be solved via standard techniques, conversely we can iterate the reduction procedure. We show that this iterative procedure terminates when either a trivial problem is encountered or the last reduced order problem is nonsingular.

1 Introduction

In this paper we consider the finite horizon, state feedback H^∞ control problem for linear time-varying systems.

This problem has been solved in the *nonsingular case* (in other words when the direct feedthrough matrix $D(\cdot)$ between the control input and the controlled output is full column rank), see for example [8], [6] and [5].

However, as often it happens in real situations, the above-mentioned $D(\cdot)$ matrix *is not* full column rank; in that case the problem is said to be *singular*. A preliminary result concerning the singular case can be found in [3], where the *full information* problem has been solved using the theory developed in the differential game framework [2].

Here we deal with the more important *state feedback* singular H^∞ problem. First we shall show that, if the matrix $D(\cdot)$ is identically zero and the number of control inputs equals the order of the system, the solution of the H^∞ problem is obtained via the simple inversion of the input matrix; in the general case of a nonzero $D(\cdot)$, our main result consists in proving that the original singular H^∞ problem is *equivalent* to another H^∞ problem related to a *reduced order* system. If this problem turns out to be nonsingular, it can be solved via standard techniques; conversely we can iterate the reduction procedure. This iterative algorithm terminates when either a trivial problem is encountered (one of the reduced systems is such that the feedthrough matrix is zero and the input matrix is square) or one of the reduced order problems is nonsingular.

An alternative approach could be that of modifing the controlled variable of the original system by introducing a sufficiently small fictitious parameter ε, so as to render the matrix $\boldsymbol{D}(\,\cdot\,)$ full column rank; however this way, as known, leads to a bad conditioned (stiff) differential Riccati equation; conversely the differential Riccati equation associated to the last reduced system in our approach does not suffer of this drawback, being related to a full column rank feedthrough matrix.

This work extends to the time-varying setting the results contained in the paper by Stoorvogel [7], where the singular H^∞ control problem for *time-invariant* systems has been solved by means of an elegant decomposition of the state space involving the concept of strongly controllable subspace. Such decomposition is not possible in the time-varying setting; therefore in the last case the machinery needed to solve the problem is based on a different approach.

The paper is organized as follows: the singular H^∞ control problem we deal with is stated precisely in Section 2; our machinery is developed in Section 3, while in Section 4 the procedure to solve the singular H^∞ problem is detailed; finally in Section 5 an example of application of our methodology is provided.

For the sake of conciseness we have omitted the proofs of the theorems which are long and technical; the interested reader is referred to [1].

2 Notation, Statement of the Problem and Preliminary Results

Throughout the paper matrices and vectors are denoted by boldface capital and boldface small letters respectively; operators are denoted by capital calligraphic letters.

In the sequel we shall refer to the following linear time-varying system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}(t)\boldsymbol{x}(t) + \boldsymbol{B}(t)\boldsymbol{u}(t) + \boldsymbol{H}(t)\boldsymbol{w}(t)\,, \tag{1a}$$

$$\boldsymbol{z}(t) = \boldsymbol{C}(t)\boldsymbol{x}(t) + \boldsymbol{D}(t)\boldsymbol{u}(t) \quad \boldsymbol{x}(t_0) = 0 \tag{1b}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^n$ is the system state, $\boldsymbol{u}(t) \in \mathbb{R}^m$ is the control input, $\boldsymbol{w}(t) \in \mathbb{R}^r$ is an exogenus disturbance input and $\boldsymbol{z}(t) \in \mathbb{R}^q$ is the controlled variable. The system matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{H}$, $\boldsymbol{C}$, $\boldsymbol{D}$, are bounded, piecewise continuously differentiable functions of $t \in \Omega := [t_0, t_f]$. The matrices $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ are assumed to have constant rank in Ω.

Given $\boldsymbol{R} : \Omega \to \mathbb{R}^{n\times n}$, we write $\boldsymbol{R} > 0$ (≥ 0) to mean that $\boldsymbol{R}$ is positive definite (semidefinite), i. e. that there exists $\alpha > 0$ such that for all $\boldsymbol{v} \in \mathbb{R}^n$ and for all $t \in \Omega$

$$\boldsymbol{v}^T\boldsymbol{R}(t)\boldsymbol{v} \geq \alpha\boldsymbol{v}^T\boldsymbol{v} \qquad (\boldsymbol{v}^T\boldsymbol{R}(t)\boldsymbol{v} \geq 0)\,.$$

In order to avoid cumbersome notation, throughout the paper we will often omit the time argument.

We recall that a controller in the form

$$\mathcal{K} : L^2(\Omega) \to L^2(\Omega), \boldsymbol{x} \to \boldsymbol{u}\,, \tag{2}$$

is said to be a linear *state feedback* controller, while

$$\mathcal{K} : L^2(\Omega) \times L^2(\Omega) \to L^2(\Omega), (\boldsymbol{x}, \boldsymbol{w}) \to \boldsymbol{u}\,. \tag{3}$$

is said to be a linear *full information* controller.

If $w \in L^2(\Omega)$, $\|w\|_{Q,2}$ denotes the norm of w in $L^2(\Omega)$ weighted by the bounded, positive definite matrix valued function $Q(\cdot)$; we let $\|w\|_2 := \|w\|_{I,2}$, where I is the constant identity matrix.

Given a linear operator $\mathcal{G}$ which maps $L^2(\Omega)$ into $L^2(\Omega)$, $\|\mathcal{G}\|$ denotes the norm of $\mathcal{G}$ induced by the usual norm in $L^2(\Omega)$, i.e.

$$\|\mathcal{G}\| = \sup_{w\in L^2(\Omega)-0} \frac{\|\mathcal{G}w\|_2}{\|w\|_2} = \max_{\|w\|_2=1} \|\mathcal{G}w\|_2 \tag{4}$$

Finally the identical and the zero operators are denoted by $\mathcal{I}$ and 0 respectively.

The goal of this paper is the solution of the following problem

State Feedback (SF) Problem Let us denote by $\mathcal{T}_{SF}$ the closed loop operator, mapping w to z, obtained when a *state feedback* controller is applied to system (1). Then, given $\gamma > 0$, find a state feedback controller such that

$$\|\mathcal{T}_{SF}\| < \gamma .$$

■

The SF Problem has been solved in the nonsingular case, that is when D is full column rank; the proof of the following theorem can be found in [5].

Theorem 1 *Assume that D is full column rank. Then the SF Problem admits a solution* if and only if *the Riccati differential equation*

$$-\dot{P} = PA + A^TP + \frac{1}{\gamma^2}PHH^TP + C^TC - (PB + C^TD)(D^TD)^{-1}(B^TP + D^TC)$$
$$P(t_f) = O \tag{5}$$

admits a positive semidefinite solution in Ω. In this case a controller which solves the SF Problem is given by

$$u = -Fx , \tag{6}$$

where († denotes the Moore-Penrose pseudoinversion)

$$F = D^{\dagger}C + (D^TD)^{-1}B^TP . \tag{7}$$

Note that, as in the time-invariant case, the state feedback controller can be chosen, without loss of generality, memoryless.

Our goal in this paper is that of *solving the SF Problem in the more general case in which D is not restricted to be full column rank.*

3 Machinery

In this section we shall provide a number of theorems regarding the H^∞ control problem for singular systems; moreover we shall introduce some input and state space transformations. Such results will be used in Section 4 to solve the SF Problem.

3.1 Main Technical Results

Theorem 2 *Consider system (1) with* $B = I$ *and* $D = O$*; then for any fixed* $\gamma > 0$ *there exists a sufficiently small* $\mu > 0$ *such that the controller*

$$u = -\frac{1}{\mu}x \tag{8}$$

solves the SF Problem.

Proof. See [1]. ■

Now assume that $B = (\,B_1 \quad O\,)$, partition conformably $D = (\,D_1 \quad D_2\,)$ and consider the system

$$\begin{aligned} \dot{x} &= Ax + B_1u_1 + Hw && \text{(9a)} \\ \overline{z} &= (I - D_2D_2^\dagger)Cx + (I - D_2D_2^\dagger)D_1u_1 && \text{(9b)} \end{aligned}$$

Problem 1 Let us denote by $\mathcal{P}$ the closed loop operator mapping w to $\overline{z}$ obtained when a state feedback controller is applied to system (9). Then, given $\gamma > 0$, find a state feedback controller such that

$$\|\mathcal{P}\| < \gamma$$

Theorem 3 *Consider system (1) with* $B = (\,B_1 \quad O\,)$ *and* $D = (\,D_1 \quad D_2\,)$*; then the SF Problem admits a solution* if and only if *Problem 1 admits a solution. Moreover, if the controller* $\overline{\mathcal{K}}$ *solves Problem 1, a solution for the SF Problem is given by the controller*

$$\mathcal{K} = \begin{pmatrix} \overline{\mathcal{K}} \\ -D_2^\dagger(C\mathcal{I} + D_1\overline{\mathcal{K}}) \end{pmatrix}. \tag{10}$$

Proof. See [1] ■

Assume now that

$$B = \begin{pmatrix} B_{11} & O \\ B_{21} & I \end{pmatrix} \qquad D = (\,D_1 \quad O\,)\,,$$

partition conformably

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \qquad H = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} \qquad C = (\,C_1 \quad C_2\,)$$

and consider the two systems

$$\begin{aligned} \dot{x}_1 &= A_{11}x_1 + A_{12}x_2 + B_{11}u_1 + H_1w && \text{(11a)} \\ \dot{x}_2 &= A_{21}x_1 + A_{22}x_2 + B_{21}u_1 + u_2 + H_2w && \text{(11b)} \\ z &= C_1x_1 + C_2x_2 + D_1u_1\,, && \text{(11c)} \end{aligned}$$

$$\begin{aligned} \dot{x}_1 &= A_{11}x_1 + F_1v + H_1w && \text{(12a)} \\ z &= C_1x_1 + F_2v\,, && \text{(12b)} \end{aligned}$$

where

$$\begin{aligned} F_1 &= (\,A_{12} \quad B_{11}\,) && \text{(13)} \\ F_2 &= (\,C_2 \quad D_1\,)\,. && \text{(14)} \end{aligned}$$

Let us define the following related problems.

Problem 2 Denote by $\mathcal{R}_1$ the closed loop operator mapping w to z obtained when a state feedback controller is applied to system (11). Then, given $\gamma > 0$, find a state feedback controller such that

$$\|\mathcal{R}_1\| < \gamma \tag{15}$$

Problem 3 Denote by $\mathcal{R}_2$ the closed loop operator mapping w to z obtained when a state feedback controller is applied to system (12). Then, given $\gamma > 0$, find a state feedback controller such that

$$\|\mathcal{R}_2\| < \gamma \tag{16}$$

Theorem 4 *Problem 2 admits a solution* if and only if *Problem 3 admits a solution.*

Proof. See [1] ∎

From Theorem 4 one can derive the following corollary.

Corollary 1 *Assume that the control law*

$$v = K_s x_1 = \begin{pmatrix} K_{s1} \\ K_{s2} \end{pmatrix} x_1 , \tag{17}$$

where the partition of the matrix K_s is consistent with that of F_1, solves Problem 3. Then for a sufficiently small $\mu > 0$ the state feedback control law

$$\begin{aligned} u_1 &= K_{s2} x_1 && (18a) \\ u_2 &= -\frac{1}{\mu} \tilde{x}_2 , && (18b) \end{aligned}$$

where $\tilde{x}_2 = x_2 - K_{s1} x_1$, solves Problem 2.

Proof. See [1] ∎

Remark 1 Corollary 1 is applied as follows: for a given $\mu > 0$ (for example $\mu = 1$) the norm of the input-output operator associated to the closed loop system, consisting of system (11) and controller (18), is evaluated. If the norm satisfies (15) then the value of μ is ok; otherwise the value of μ is reduced until the closed loop system satisfies (15). We recall that, given the linear operator $\mathcal{G}$ associated to a linear system, the norm (4) can be numerically evaluated solving a certain Riccati differential equation (see [9]).

3.2 Three Useful Transformations

Transformation 1

Given system (1), where $\operatorname{rank}(B) = p < m$ *(that is B is not full column rank), there exists an orthonormal input trasformation*

$$u = V\tilde{u} = V \begin{pmatrix} u_1 \\ \begin{pmatrix} u_2 \\ u_3 \end{pmatrix} \end{pmatrix} ,$$

which puts the system into the form

$$\begin{aligned} \dot{x} &= Ax + B_1 u_1 + Hw && (19a) \\ z &= Cx + D_1 u_1 + D_2 u_2 , && (19b) \end{aligned}$$

where the matrices $B_1(t) \in \mathbb{R}^{n \times p}$ and D_2 are full column rank.

Proof. Let

$$p = \operatorname{rank}(\boldsymbol{B}) < m \, . \tag{20}$$

In this case there exists an orthonormal trasformation $\boldsymbol{V}_1$ such that

$$\boldsymbol{B}\boldsymbol{V}_1 = \begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{O} \end{pmatrix} , \tag{21}$$

where $\boldsymbol{B}_1$ is full column rank.

Let

$$\boldsymbol{D}\boldsymbol{V}_1 = \begin{pmatrix} \boldsymbol{D}_1 & \overline{\boldsymbol{D}}_2 \end{pmatrix} , \tag{22}$$

with $\boldsymbol{D}_1(t) \in \mathbb{R}^{q\times p}$ and $\boldsymbol{D}_2(t) \in \mathbb{R}^{q\times(m-p)}$.

Again we can find an orthonormal transformation $\boldsymbol{V}_2$ such that

$$\overline{\boldsymbol{D}}_2\boldsymbol{V}_2 = \begin{pmatrix} \boldsymbol{D}_2 & \boldsymbol{O} \end{pmatrix} \tag{23}$$

where $\boldsymbol{D}_2$ is full column rank.

Now, letting

$$\boldsymbol{V} = \boldsymbol{V}_1 \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{V}_2 \end{pmatrix} , \tag{24}$$

the proof follows. ■

Transformation 2 *Given system (1), where* $\boldsymbol{B}$ *is full column rank and* $\operatorname{rank}(\boldsymbol{D}) = p < m$, *there exists an orthonormal input trasformation*

$$\boldsymbol{u} = \boldsymbol{V}\tilde{\boldsymbol{u}} = \boldsymbol{V} \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{pmatrix} ,$$

which puts the system into the form

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}_1\boldsymbol{u}_1 + \boldsymbol{B}_2\boldsymbol{u}_2 + \boldsymbol{H}\boldsymbol{w} && \text{(25a)} \\ \boldsymbol{z} &= \boldsymbol{C}\boldsymbol{x} + \boldsymbol{D}_1\boldsymbol{u}_1 \, , && \text{(25b)} \end{aligned}$$

where the matrices $\boldsymbol{D}_1(t) \in \mathbb{R}^{q\times p}$, $\boldsymbol{B}_1(t) \in \mathbb{R}^{n\times p}$ *and* $\boldsymbol{B}_2(t) \in \mathbb{R}^{n\times(m-p)}$ *are full column rank.*

Proof. Let

$$p = \operatorname{rank}(\boldsymbol{D}) < m \, . \tag{26}$$

In this case there exist an orthonormal trasformation $\boldsymbol{V}$ such that

$$\boldsymbol{D}\boldsymbol{V} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{O} \end{pmatrix} , \tag{27}$$

where $\boldsymbol{D}_1$ is full column rank.

Letting

$$\begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_2 \end{pmatrix} = \boldsymbol{B}\boldsymbol{V} \, , \tag{28}$$

the proof follows. ■

Transformation 3 *(Butman transformation [4])*
Consider system (25) with $\boldsymbol{B}_1(t) \in \mathbb{R}^{n\times p}$, $\boldsymbol{B}_2(t) \in \mathbb{R}^{n\times(m-p)}$ *and* $\boldsymbol{D}_1(t) \in \mathbb{R}^{q\times p}$ *full column rank; assume* $0 < m - p < n$. *Then there exists a state space transformation which puts system (25) into the form (11).*

Proof. Let $E(t) \in \mathbb{R}^{n\times(n-m+p)}$ be such that

$$E^T B_2 = O \qquad E^T E = I\,. \tag{29}$$

By means of the state space transformation

$$x = \begin{pmatrix} E & B_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{30}$$

system (25) can be rewritten as (11) where

$$\begin{aligned}
A_{11} &= \dot{E}^T E + E^T A E && \text{(31a)}\\
A_{12} &= \dot{E}^T B_2 + E^T A B_2 && \text{(31b)}\\
A_{21} &= B_2^{\dagger}(AE - \dot{E}) && \text{(31c)}\\
A_{22} &= B_2^{\dagger}(AB_2 - \dot{B}_2) && \text{(31d)}\\
B_{11} &= E^T B_1 && \text{(31e)}\\
B_{21} &= B_2^{\dagger} B_1 && \text{(31f)}\\
H_1 &= E^T H && \text{(31g)}\\
H_2 &= B_2^{\dagger} H && \text{(31h)}\\
C_1 &= CE && \text{(31i)}\\
C_2 &= CB_2\,. && \text{(31j)}
\end{aligned}$$

■

4 Solution of the SF Problem

Now we come back to the solution of the SF Problem. In this section we shall assume that $B(t) \in \mathbb{R}^{n\times m}$ is full column rank. If this is not the case, by applying Transformation 1 and using Theorem 3, we can restate the problem referring to a system with a full column rank B matrix.

Now if $\text{rank}(D) =: p = m$ (that is D is full column rank) the SF Problem turns out to be a regular H^∞ problem which can be solved via Theorem 1.

In the other cases ($p < m$) we propose the following procedure.

Procedure 1

Step 1 Using Transformation 2, system (1) can be put into the form (25), which we rewrite here for convenience

$$\begin{aligned}
\dot{x} &= Ax + B_1 u_1 + B_2 u_2 + Hw && (25a)\\
z &= Cx + D_1 u_1\,, && (25b)
\end{aligned}$$

with $D_1(t) \in \mathbb{R}^{q\times p}$, $B_1(t) \in \mathbb{R}^{n\times p}$ and $B_2(t) \in \mathbb{R}^{n\times(m-p)}$ full column rank matrices.

Step 2 Three different cases have to be discussed.

Case 1 $p = 0$ *and* $m = n$. In this case the dimension of the vector $\boldsymbol{u}_1$ is zero, while the dimension of the vector $\boldsymbol{u}_2$ is n; therefore the matrix $\boldsymbol{B}_2$ is invertible. The SF Problem can be solved with the aid of Theorem 2; indeed a controller solving the SF Problem is

$$\boldsymbol{u}_2 = -\frac{1}{\mu}\boldsymbol{B}_2^{-1}\boldsymbol{x}\,,$$

where $\mu > 0$ is sufficiently small.

Case 2 $p = 0$ *and* $m < n$. In this case using Transformation 3 we put system (25) into the form (11); then applying Theorem 4, the SF Problem is re-stated, via Problem 3, for the reduced order system (12)

$$\begin{aligned} \dot{\boldsymbol{x}}_1 &= \boldsymbol{A}_{11}\boldsymbol{x}_1 + \boldsymbol{F}_1\boldsymbol{v} + \boldsymbol{H}_1\boldsymbol{w} && (12a)\\ \boldsymbol{z} &= \boldsymbol{C}_1\boldsymbol{x}_1 + \boldsymbol{F}_2\boldsymbol{v}\,. && (12b) \end{aligned}$$

If this problem is regular it can be solved via Theorem 1, conversely redefine system (1) as system (12) and *go to* Step 1.

Case 3 $0 < p < m$. This situation can be treated in the same way of Case 2.

This iterative procedure terminates when either a trivial problem is encountered (Case 1 of Step 2) or one of the reduced order problems is nonsingular .

5 An Example

Consider system (1) with $t \in [0.2, 1.2]$ and

$$\boldsymbol{A} = \begin{pmatrix} 0 & e^{-t} & 0\\ 0 & \cos t & 1\\ 0 & -\sin t & 1 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 2 & -1\\ \frac{2}{\sin t} & -\frac{1}{\sin t}\\ -\frac{1+e^{-t}}{\sin t} & \frac{2+e^{-t}}{\sin t} \end{pmatrix}, \quad \boldsymbol{H} = \begin{pmatrix} 0\\ \sin t\\ 1 \end{pmatrix}; \qquad (32a)$$

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & 1\\ 1 & 1 & 0 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} e^{-t} & -e^{-t}\\ \cos t & -\cos t \end{pmatrix}, \quad \gamma = 1\,. \qquad (32b)$$

Now, since $p := \operatorname{rank}(\boldsymbol{D}) = 1 < m := \operatorname{rank}(\boldsymbol{B}) = 2$, we apply Procedure 1.

Step 1 We apply the orthonormal input Transformation 2 via the matrix

$$\boldsymbol{V} = \begin{pmatrix} \cos t & \sin t\\ -\cos t & \sin t \end{pmatrix}.$$

which puts the system into the form (25) with

$$\boldsymbol{B}_1 = \begin{pmatrix} 3\cos t\\ 3\frac{\cos t}{\sin t}\\ -(3+2e^{-t})\frac{\cos t}{\sin t} \end{pmatrix} \quad \boldsymbol{B}_2 = \begin{pmatrix} \sin t\\ 1\\ 1 \end{pmatrix}. \quad \boldsymbol{D}_1 = \begin{pmatrix} 2e^{-t}\cos t\\ 2\cos^2(t) \end{pmatrix}$$

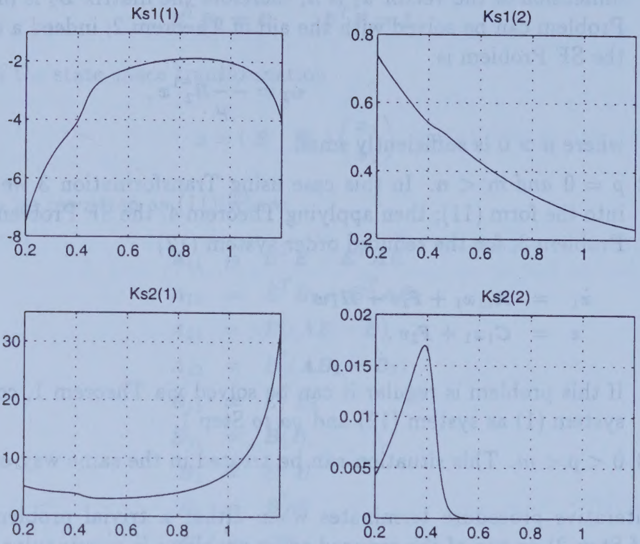


Figure 1: Gains of the optimal controller

Step 2 We apply the state space Transformation 3 via the matrix $\begin{pmatrix} \boldsymbol{E} & \boldsymbol{B}_2 \end{pmatrix}$ with

$$\boldsymbol{E} = \begin{pmatrix} 0 & \frac{1}{\sqrt{1+0.5\sin^2 t}} \\ \frac{1}{\sqrt{2}} & -\frac{\sin t}{2\sqrt{1+0.5\sin^2 t}} \\ -\frac{1}{\sqrt{2}} & -\frac{\sin t}{2\sqrt{1+0.5\sin^2 t}} \end{pmatrix}$$

which puts the system (25) into the form (11). Now applying Theorem 4, the SF Problem is re-stated, via Problem 3, for the reduced order system (12); note that the matrix (see (14) and (31j))

$$\boldsymbol{F}_2 = \begin{pmatrix} \sin t + 1 & 2e^{-t}\cos t \\ \sin t + 1 & 2\cos^2(t) \end{pmatrix}$$

is full column rank for all $t \in [0.2, 1.2]$, therefore the reduced order problem is *regular*. According to the notation (17), the optimal control law for the reduced problem, obtained by solving via numerical software the differential Riccati equation (5), is in the form

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{K}_{s1}(1) & \boldsymbol{K}_{s1}(2) \\ \boldsymbol{K}_{s2}(1) & \boldsymbol{K}_{s2}(2) \end{pmatrix} \boldsymbol{x}_1 \,. \tag{33}$$

The time-varying gains of the optimal controller for the reduced order system (12) are depicted in Figure 1.

The optimal control law for the original system (1) with matrices (32) is given by (18) with $\mu = 0.5$.

6 Conclusions

In this paper we have considered the H^∞ control problem for linear time-varying systems when the feedthrough matrix $\boldsymbol{D}$ is not full column rank, the so-called singular H^∞ problem. If $\boldsymbol{D}$ is zero and the number of control inputs equals the order of the system, the problem is solved via simple inversion of the input matrix; when $\boldsymbol{D}$ is nonzero we show that the original problem can be restated for a reduced order system. If this new problem is nonsingular, a solution can be found via standard techniques, while if it is still singular we can apply for a second time the reduction procedure. The algorithm terminates when either a trivial problem is encountered or one of the reduced order problems is nonsingular.

References

[1] F. Amato, M. Mattei, A. Pironti, "Solution of the state feedback singular H^∞ control problem for linear time-varying systems", *Internal Report*, University of Naples, 1997.

[2] F. Amato, A. Pironti, "A note on singular zero-sum linear quadratic differential games," *Proceedings of the 33rd IEEE Conference on Decision and Control*, Orlando, December 1994.

[3] F. Amato, A. Pironti, "Singular finite horizon full information H^∞ Control via Reduced Order Riccati Equations," *Kybernetika*, Vol.31 (1995), pp. 599–609.

[4] S. Butman, "A method for optimizing control-free costs in systems with linear controllers," *IEEE Trans. Aut. Contr.*, (1968), pp. 554–556.

[5] D. J. N. Limebeer, B. D. O. Anderson, P. P. Khargonekar, M. Green, "A game theoretic approach to H^∞ control for time-varying systems," *SIAM J. Contr. Opt.*, vol. 30 (1992), pp. 262–283.

[6] R. Ravi, K. M. Nagpal, P. P. Khargonekar, "H^∞ control of linear time-varying systems: a state space approach," *SIAM J. Contr. Opt.*, vol. 29 (1991), pp. 1394–1413.

[7] A. A. Stoorvogel, H. Trentelman, "The quadratic matrix inequality in singular H^∞ control with state feedback," *SIAM J. Contr. Opt.*, vol. 28 (1990), pp. 1190–1208.

[8] G. Tadmor, "Worst-case design in time domain: the maximum principle and the standard H^∞ problem," *MCSS*, vol. 3 (1990), pp. 301–324.

[9] G. Tadmor, "Input/Output norms in general linear systems," *Int. J. Contr.*, vol. 51 (1990), pp. 911–921.

Fundamental Design Constraints in Linear Filtering and Control

Graham C. Goodwin   Centre for Integrated Dynamics and Control
The University of Newcastle, NSW 2308, Australia

Abstract

Design constraints are central to many engineering problems. Examples are the Cramer Rao inequality in point estimation theory, the second Law of Thermodynamics and Shannon's Bandwidth Error result in Communications. In this paper, we review design constraints in linear filtering. We show that there are fundamental and unavoidable limitations on the achievable performance. The corresponding results in feedback control are also briefly discussed.

1 Introduction

Fundamental design constraints lie at the core of many engineering problems. For example, it is well known that the Second Law of Thermodynamics precludes the design of certain machines which would otherwise be highly attractive.

Here we address the problem of estimating unmeasured variables from certain measured variables. For simplicity, we restrict attention to linear time-invariant signal models. General background on linear filters may be found in [14], [15], [16]. However, it is to be expected that related constraints will apply to time-varying and/or nonlinear problems. This is a topoic of current research activity.

A well known result of the type we will discuss here is the Bode gain-phase relationship [1]. This result states that there is a one-to-one relationship between the gain and phase response of a linear minimum phase system.

The topic of fundamental design constraints has been extensively studied in the area of linear control system design - the interested reader is referred to [2], [3], [4], [5], [6], [7]. A comprehensive treatment of these and related results is contained in [8].

Fundamental Design constraints in filtering have also been discussed in recent literature. For example, [9] discusses fundamental constraints on filter sensitivity functions for unbiased estimators (UE's) [13]. These results are extended to a broader class of filters, called Bounded Error Estimators (BEE's), in [10] or State Error Estimators (SEE's) ([11]). Limiting L_2 performance of filters is discussed in [17].

2 Signal Model

We will assume that the signal model is linear time-invariant and is given by:

$$\overset{\circ}{x} = Ax + B_1 u + B_2 \omega ; x(0) = x_0 \tag{1}$$

$$y = Cx + v; \; y^0 = Cx \tag{2}$$

$$z = Dx \tag{3}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $\omega \in \mathbb{R}^s$, $y \in \mathbb{R}^p$ and $z \in \mathbb{R}^q$. We assume that u is the measured input (possibly zero) and y is the measured output. We also assume that $(C; A)$ is detectable, but $(A; B_1, B_2)$is not necessarily stabilizable. Our problem is to estimate z via a linear filter driven by y and possibly u. We make no particular assumptions about x_0, ω or v. However, a major issue will be the sensitivity of the estimation error to x_0, ω, v.

3 A General Class of Filters

Most contemporary filters have a state dimension equal to that of the system. Here we will avoid making this assumption. Instead, we consider a very general class of filters defined as follows:

$$\overset{\circ}{x}_f = A_f x_f + B_f u + E_f y; \, x_f(0) = x_{fo} \tag{4}$$

$$\hat{x} = C_f x_f, \; \hat{z} = D\hat{x}, \; \hat{y} = C\hat{x} \tag{5}$$

where $x_f \in \mathbb{R}^{n_f}$. Note that the above filter is strictly proper. It would be straightforward to also include direct feedthrough terms but we will avoid this possibility for simplicity. We also assume that the filter is stable, i.e. A_f has all its eigenvalues in the strict left half plane. We need to further restrict the filter given in (4), (5) so that it has interesting properties. For example, nothing has been said, up to this point, about the quality of the estimates $\hat{x}$ or $\hat{z}$. We therefore restrict attention to stable error estimators (SEE). These are defined as follows:

Definition 1 *A filter is said to be a Stable Error Estimator, if for $\omega = 0, \; v = 0, \; u = 0$ then $\tilde{x}(t) = x(t) - \hat{x}(t) \to 0$ as $t \to \infty$.*

□

Note that for $w = 0, \; v = 0, \; u = 0$, the estimation error satisfies

$$\dot{x}_e = A_e x_e \tag{6}$$

$$\tilde{x} = C_e x_e \tag{7}$$

where

$$x_e = \begin{pmatrix} x \\ x_f \end{pmatrix}; \; A_e = \begin{bmatrix} A & 0 \\ E_f C & A_f \end{bmatrix} \tag{8}$$

$$C_e = [I \quad , -C_f] \tag{9}$$

We then have the following result ([11]):

Lemma 2 *A stable error state estimator exists if and only if $(C, \; A)$ is detectable.*

□

Proof.

If: This is a well known condition to design a stable error observer of dimension $n_f = n$

Only if: Recall that the pair (C, A) is detectable if and only $Ax = \lambda x$ for $Cx = 0$, $\operatorname{Re}\lambda \geq 0$ implies $x = 0$. Hence assume that there exists an $x_0 \neq 0$ such that $Ax_0 = \lambda x_0$ and $Cx_0 = 0$, then

$$A_e \begin{bmatrix} x_0 \\ 0 \end{bmatrix} = \begin{bmatrix} A & 0 \\ E_f C & A_f \end{bmatrix} \begin{bmatrix} x_0 \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \tag{10}$$

and

$$C_e \begin{bmatrix} x_0 \\ 0 \end{bmatrix} = x_0 \tag{11}$$

From (10), the solution to (6), (7) for initial state $x_e(0) = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$ is

$$x_e(t) = e^{A_s t} \begin{pmatrix} x_0 \\ 0 \end{pmatrix} = \begin{pmatrix} e^{\lambda t} x_0 \\ 0 \end{pmatrix}$$

$$\tilde{x}(t) = e^{\lambda t} x_0;\ t \geq 0 \tag{12}$$

Hence the stable error property necessarily implies $\operatorname{Re}\lambda < 0$; i.e. the system is detectable.

$\nabla\nabla\nabla$

4 n-Dimensional Observers

A particular case of (4), (5), occurs for $n_f = n$ and when the following choices are made:

$$A_f = A - JC \text{ is stable} \tag{13}$$

$$B_f = B \tag{14}$$

$$E_f = J \tag{15}$$

In this case, the estimation error satisfies

$$\overset{\circ}{\tilde{x}} = (A - JC)\tilde{x} \tag{16}$$

The estimator given in (13) to (16) has an additional property, namely that of unbiasedness.

Definition 3 *A stable error state estimator is said to be unbiased if the transfer function from u to $\hat{x}$ is the same as the transfer function linking u to x.*

□

Note that not all commonly used estimators have the unbiasedness property. For example, robust estimators based on H^∞ designs are often biased.

Another interesting property of Kalman filters is that they have a feedback or *innovations* structure. Specifically, if we define

$$\nu = y - C\hat{x} \tag{17}$$

then the estimator can be written

$$\overset{\circ}{\hat{x}} = A\hat{x} + Bu + J\nu \tag{18}$$

Equations (17), (18) can also be expressed using Laplace Transforms, with $\phi(s) = (sI - A)$. as

$$\hat{x}(s) = \phi(s)^{-1}[Bu(s) + J\nu(s)] \tag{19}$$

$$\nu(s) = y(s) - C\hat{x}(s) \tag{20}$$

Actually, this feedback form holds more generally for all SEE's as discussed below:

5 All Stable Error Estimators

We will temporarily suppress the measured input u since this plays no role in our present discussion. We then state without proof (see [11] for details), the result that all SEE's can be rearranged into the form shown in Figure 1 or equivalently into the error form given in Figure 2 where $\phi(s) = (sI - A)$ and where $K(s)$ necessarily gives internally stability of the loop.

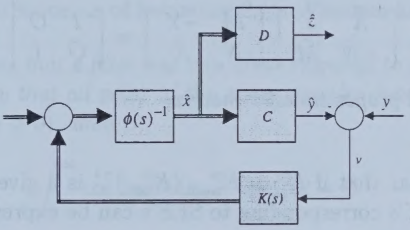


Figure 1:

Figures 1 and 2 are reminiscent of feedback control structures and allow control insights to be used for filter design.

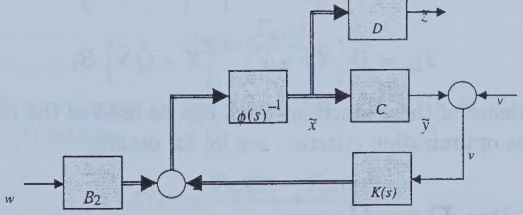


Figure 2:

Two differences arise which distinguish filtering from control. These are:

(i) $K(s)$ need not be a proper transfer function in filtering

(ii) $K(s)$ is $n \times p$, ie. there are usually extra degrees of freedom in filtering than in control.

In those cases where we restrict $K(s)$ to be proper, then we can easily obtain a characterization of all SEE's using the standard parameterization of all proper stabilizing controllers.

Theorem 4 *The class of all stable error filters in which $K(s)$ is proper is parameterized by*

$$K = \tilde{U}^{-1}\tilde{V} = VU^{-1}$$

where

$$\begin{aligned} \tilde{U} &= \tilde{X} + \tilde{Q}\tilde{N} \\ \tilde{V} &= \tilde{Y} - \tilde{Q}\tilde{M} \\ U &= X + N\tilde{Q} \\ V &= Y - M\tilde{Q} \end{aligned}$$

where

$$C\phi(s)^{-1} = NM^{-1} = \tilde{M}^{-1}\tilde{N}$$

$$\begin{bmatrix} \tilde{X} & \tilde{Y} \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -Y \\ N & X \end{bmatrix} = \begin{bmatrix} I & O \\ O & I \end{bmatrix}$$

and $\tilde{Q}$ is any stable and proper transfer matrix

□

We note, in particular that if $K^\circ = K^\circ_{num}\left(K^\circ_{den}\right)^{-1}$ is a given right mfd for some K giving a SEE then all K's corresponding to SEE's can be expressed as:

$$K = \left(K^\circ_{num} - \phi\tilde{Q}\right)\left(K^\circ_{den} + C\tilde{Q}\right)^{-1} \tag{21}$$

Using the parameterization given in Theorem 4, it can be easily seen that $T_{\tilde{z}v}$ and $T_{\tilde{z}\omega}$ are affine functions of $\tilde{Q}$, i.e.

$$T_{\tilde{z}v} = -D\left[\tilde{X}\phi + \tilde{Y}\right]^{-1}\left[\tilde{Y} - \tilde{Q}\tilde{M}\right]$$

$$T_{\tilde{z}\omega} = D\left[\tilde{X}\phi + \tilde{Y}\right]^{-1}\left[\tilde{X} + \tilde{Q}\tilde{N}\right]B_2$$

The simple dependence of these functions on $\tilde{Q}$ can be used as the basis of optimal filter design with various optimization criteria - see [8] for details.

6 Sensitivity Functions

The feedback structure of Figures 1 and 2 motivates the following transfer functions

$$T_{\tilde{z}y} = D[\phi + KC]^{-1}K \triangleq T_2 \tag{22}$$

$$T_{\tilde{z}v} = -D[\phi + KC]^{-1}K \tag{23}$$

$$T_{\tilde{z}\omega} = D[\phi + KC]^{-1}B_2 \tag{24}$$

For future reference, we also define the following open-loop transfer functions

$$T_{y\omega} = C\phi^{-1}B_2 \triangleq G_2 \tag{25}$$

$$T_{z\omega} = D\phi^{-1}B_2 \triangleq G_4 \tag{26}$$

We also note that, in terms of these quantities

$$T_{\tilde{z}\omega} = (G_4 - T_2G_2) \tag{27}$$

$$T_{\tilde{z}v} = -T_2 \tag{28}$$

In the sequel we will restrict attention to the scalar case where ω, v, y, z are all scalar quantities. However, extensions to the multivariable case are possible as in control - see for example, [6], [7].

7 Time Domain Constraints

It is possible to use (23), (24) directly to obtain various time domain constraints satisfied by the error response. As an illustration, say we wish to design a filter which has zero error to a sinusoidal disturbance, ω, of frequency $2\pi f_0$. We then have the following result.

Lemma 5 *Let us assume that a filter has zero error response to a sinusoidal disturbance of frequency $2\pi f_0$ and say that all poles of the error transfer functions $T_{\tilde{z}v}$ and $T_{\tilde{z}\omega}$ have real parts less than $-\alpha(\alpha > 0)$, then*

$$\int_0^\infty \cos 2\pi f_0 t\, [\tilde{z}_{ss} - \tilde{z}(t)]\, dt = 0 \tag{29}$$

$$\int_0^\infty \sin 2\pi f_0 t\, [\tilde{z}_{ss} - \tilde{z}(t)]\, dt = \frac{T_{\tilde{z}\omega}(0)}{2\pi f_0} \tag{30}$$

where $\tilde{z}(t)$ is the response to a unit step disturbance in $\omega(t)$ at $t = 0$ and $\tilde{z}_{ss} = \lim_{t\to\infty} \tilde{z}(t)$. □

Proof. By definition, the Laplace Transform of the step response of the filter satisfies

$$\tilde{z}(t) = \frac{T_{\tilde{z}\omega}(s)}{s} \tag{31}$$

in the region of the convergence of the transform. Let $\varepsilon(t) = \tilde{z}_{ss} - \tilde{z}(t)$. Then the transform, $E(s)$, of $\varepsilon(t)$ satisfies

$$E(s) = \frac{T_{\tilde{z}\omega}(0) - T_{\tilde{z}\omega}(s)}{s}$$

Clearly $E(s)$ does not have a pole at $s = 0$. Hence $s = \pm j2\pi f_0$ lies in the region of convergence of $E(s)$. Thus

$$\begin{aligned} E(s) &= \int_0^\infty e^{-st}\varepsilon(t)dt \text{ at } s = \pm j2\pi f_0 \\ &= \frac{T_{\tilde{z}\omega}(0) - T_{\tilde{z}\omega}(s)}{s} \end{aligned}$$

Now $T_{\tilde{z}\omega}(\pm j2\pi f_0) = 0$, therefore

$$\int_0^\infty e^{\pm j2\pi f_0 t}\varepsilon(t)dt = \frac{T_{\tilde{z}\omega}(0)}{\mp j2\pi f_0}$$

The result follows on using the fact that

$$\cos 2\pi f_0 t = \frac{1}{2}\left[e^{j2\pi f_0 t} + e^{-j2\pi f_0 t}\right]$$

$$\sin 2\pi f_0 t = \frac{1}{2j}\left[e^{j2\pi f_0 t} - e^{-j2\pi f_0 t}\right]$$

$\nabla\nabla\nabla$

The above result provides a link between the steady state response (here zero) to a sinusoidal disturbance and the transient and steady state response to a step disturbance. This result has significant implications in design. For example, a feedback control application to a rolling mill problem is discussed in [12].

8 Frequency Domain Constraints

We next explore frequency domain constraints.

It is convenient to normalize the error sensitivities given in (23), and (24). A possible normalization is given below:

We define

$$\begin{aligned} S_f &= T_{\tilde{z}\omega}T_{z\omega}^{-1} \\ &= 1 - T_2G_2G_4^{-1} \end{aligned} \tag{32}$$

$$\begin{aligned} T_f &= -T_{\tilde{z}v}T_{y\omega}T_{z\omega}^{-1} \\ &= T_2G_2G_4^{-1} \end{aligned} \tag{33}$$

Note that we immediately have the following complementarity constraint

$$S_f + T_f = I \tag{34}$$

In general, S_f and T_f will not be stable when G_4 contains non-minimum phase zeros. We will be interested in evaluating various frequency domain constraints on the magnitudes of S_f and T_f. Hence, for this purpose, we can replace unstable poles in S_f and T_f by their reflection in the imaginary axis.

It can be easily seen that the quantities S_f and T_f satisfy the following interpolation constraints.

Theorem 6 *For any SEE, we have:*

(i) If $p \in C^+$ is a pole of G_4, *then*

$$S_f(p) = 0 \quad \text{and} \quad T_f(p) = 1$$

(ii) If $q \in C^+$ is a zero of G_2 that is not simultaneously a zero of G_4, then

$$S_f(q) = 1 \quad \text{and} \quad T_f(q) = 0$$

(iii) If $q \in C^{+}$ is a zero of G_4 that is not also a zero of T_2G_2 then $S_f(s)$ and $T_f(s)$ have a pole at $s = q$.

□

Proof.

(i) From (27), $[G_4 - T_2G_2]$ is stable. The result then follows from (32).

(ii) Follows from (33)

(iii) Immediate from (32), (33)

∇∇∇

We next show that the interpolation constraints given in Theorem 6 can be translated into frequency domain integral constraints. For convenience we define.

$$Z_S \triangleq \{s \in C^{+} : S_f(s) = 0\} \tag{35}$$

$$Z_T \triangleq \{s \in C^{+} : T_f(s) = 0\} \tag{36}$$

$$Z_Z \triangleq \{s \in C^{+} : G_4(s) = 0 \text{ and } T_2(s)G_2(s) \neq 0\} \tag{37}$$

We also introduce the notation, $I_q(H)$, for the Poisson integral: Given a proper transfer function H analytic in the ORHP and $q = \sigma_q + j\omega q \in C^{+}$,

$$I_q(H) \triangleq \int_{-\infty}^{\infty} \log|H(j\omega)| \frac{\sigma_q}{\sigma_q^2 + (\omega_q - \omega)^2} d\omega \tag{38}$$

Also, let the sets (35), (36), (37) be given by $Z_S = \{p_i,\ i = 1,\ \ldots,\ n_p\}$, $Z_T = \{q_i : i = 1, \ldots, n_q\}$ and $Z_Z = \{z_i : i = 1, \ldots, n_z\}$. We introduce the Blaschke products:

$$B_S(s) = \prod_{i=1}^{n_p} \frac{s - p_i}{s + \bar{p}_i}; \quad B_T(s) = \prod_{i=1}^{n_q} \frac{s - q_i}{s + \bar{q}_i} \tag{39}$$

$$B_Z(s) = \prod_{i=1}^{n_z} \frac{s - z_i}{s + \bar{z}_i} \tag{40}$$

We then have:

Theorem 7 *For any SEE such that S_f and T_f are proper. Then*

(i) If $q = \sigma_q + j\omega q$, $\sigma_q > 0$ is a zero of G_2 that is not also a zero of G_4, i.e. $q \in Z_T$ then

$$I_q(S_f) = \pi \log|B_S(q)^{-1}| - \pi \log|B_Z(q)^{-1}| \tag{41}$$

(ii) If $p = \sigma_p + j\omega p$, $\sigma_p > 0$ is a pole of G_4, i.e. $p \in Z_S$, then

$$I_p(T_f) = \pi \log|B_T(p)^{-1}| - \pi \log|B_Z(p)^{-1}| \tag{42}$$

□

Proof. This follows from a straightforward application of the Poisson integral formula as in [2]; see also [8].

∇∇∇

From the above result it can be inferred that in the presence of right half plane poles or zeros, then a reduction of $|S_f(j\omega)|$ or $|T_f(j\omega)|$ in one frequency range must be accompanied by an increase in another frequency range. To quantify this trade-off we note that typical design requirements are that $|S_f(j\omega)|$ be small at low frequencies whilst $|T_f(j\omega)|$ should be small at high frequencies. Thus we might require

$$|S_f(j\omega)| < \alpha_1 \quad \text{for} \quad \omega \in [0,\ \omega_1] \tag{43}$$

$$|T_f(j\omega)| < \alpha_2 \quad \text{for} \quad \omega \in [\omega_2,\ \infty] \tag{44}$$

where α_1 and α_2 are small positive numbers. Note that (44) is equivalent to the following (Using (34))

$$|S_f(j\omega)| < 1 + \alpha_2 \quad \text{for} \quad \omega \in [\omega_2,\ \infty] \tag{45}$$

Once design requirements such as (43), (44) have been established, then Theorem 7 can be used to quantify the performance in other frequency ranges. To illustrate this point, consider a frequency interval of the form $[-\omega_1,\ \omega_1]$ and define, for each complex number $s_0 = \sigma_0 + j\omega_0$, the weighted length of the interval $[-\omega_1,\ \omega_1]$ as

$$\theta_{s_0}(\omega_1) \triangleq \int_{-\omega_1}^{\omega_1} \frac{\sigma_0}{\sigma_0^2 + (\omega_0 - \omega)^2} d\omega \tag{46}$$

Under these conditions, it is easy to show that when S_f and T_f are proper and G_4 has no unstable zeros, then for each $q = \sigma_q + j\omega_q$ that is a rhp zero of G_2

$$\begin{aligned} \|S_f\|_\infty \ \geq\ & \left(\frac{1}{\alpha_1}\right)^{\theta_q(\omega_1)/(\theta_q(\omega_2)-\theta_q(\omega_1))} \\ & \times \left(\frac{1}{1+\alpha_2}\right)^{(\pi-\theta_q(\omega_2))/(\theta_q(\omega_2)-\theta_q(\omega_1))} \\ & \times \left|B_S^{-1}(q)\right|^{\pi/(\theta_q(\omega_2)-\theta_q(\omega_1))} \end{aligned} \tag{47}$$

Similarly, we can see that for each $p = \sigma_p + j\omega_p$ that is a right half plane pole of G_4,

$$\begin{aligned} \|T_f\|_\infty \ \geq\ & \left(\frac{1}{\alpha_2}\right)^{(\pi-\theta_p(\omega_2))/(\theta_p(\omega_2)-\theta_p(\omega_1))} \\ & \times \left(\frac{1}{1+\alpha_1}\right)^{\theta_p(\omega_1)/(\theta_p(\omega_2)-\theta_p(\omega_1))} \\ & \times \left|B_T^{-1}(p)\right|^{\pi/(\theta_p(\omega_2)-\theta_p(\omega_1))} \end{aligned} \tag{48}$$

These constraints quantify the achievable sensitivity functions based only on the characteristics of the signal model. The impact of these results is qualitatively similar to results in [2], [7] etc. for feedback control problems. An interesting distinction from the control case is that right-half-plane zeros of $G_4(s)$ actually weaken the constraints on filtering sensitivities. This is heuristically reasonable since a right-half-plane zero in $G_4(s)$ corresponds to a *delayed* state and hence one might expect that estimation would be easier. A similar remark applies to the constraints on a fixed lag smoothed estimate (see[[10]]).

The constraints (47), (48) also allow one to determine suitable bandwidths over which it is sensible to reduce $S_f(j\omega)$ or $T_f(j\omega)$. A full discussion is contained in [10].

As a simple example, consider the following state space model

$$A = \begin{bmatrix} 0 & 1 \\ 3 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 2 & 1 \end{bmatrix}$$

$$D = \begin{bmatrix} -2, & 1 \end{bmatrix}$$

Here we have p=1$\in Z_S$ and $q = 2 \in Z_T$ and hence $B_S^{-1}(q) = B_T^{-1}(p) = 3$.

Let b_S denote the lower bound on $\|S_f\|_\infty$ given by (47) and b_T denote the lower bound on $\|T_f\|_\infty$ given by (48). Suppose we choose $\omega_1 = 1$, $\alpha_2 = 0.5$, $\omega_2 = 100$. Then we obtain the bounds given in Figure (3) and Figure (4) on $\|S_f\|_\infty$ and $\|T_f\|_\infty$. Note these bounds hold for *any* filter design. Also shown are the results achieved by particular Kalman Filter (KF) and H_∞ designs.

As another example of the application of these results, consider the problem of estimating the angle, θ, of an inverted pendulum when the only measurements available are the force applied to the carriage and the position of the carriage. This system has a right-half-plane zero to the left of a right-half-plane pole. The constraints (47), (48) can show that a very large (100:1) sensitivity peak is unavoidable in the estimation of θ. This implies that this is an inherently difficult problem and that filtering is unlikely to be successful (independently of how the filter is designed). Full details may be found in [9].The result is physically reasonable since the estimation of θ is akin to trying to guess the angle of a broom being balanced on one's hand when one's eyes are closed; i.e. when θ is not directly observed.

9 Unbiased Estimators

So far we have not explicitly accounted for the measured input signal u since this was not relevant to the discussion of SEE's. To include considerations of a measured input we will add a stable transfer function T_1 linking $\hat{z}$ to u to the stable transfer function T_2 linking $\hat{z}$ to y; i.e. we extend the definition of $\hat{z}$ to:

$$\hat{z} = T_1 u + T_2 y \tag{49}$$

We also consider noise which comes through the same channel as u in the signal model (1), i.e. we take $B_2 = B_1$.

We recall the definition of unbiasedness given in Definition 4.1. It is readily seen that for (49) to be unbiased we require

$$D\phi^{-1}B_1 = T_1 + T_2 C \phi^{-1} B_1 \tag{50}$$

However, using (25) to (28) with $B_1 = B_2$, we now see that for unbiasedness we require that T_2 in (49) actually be $T_{\hat{z}\omega}$ for input noise ($B_1 = B_2$) i.e.

$$T_2 = T_{\hat{z}\omega} \tag{51}$$

10 Connections to Feedback Control

Finally we link state estimation to feedback control. Consider a SISO system as in (1) to (3). We will assume that the system has a deterministic input disturbance. For simplicity, we assume a step disturbance although finite sums of sinusoids can be similarly treated. In this case, we take the following partitioned form for the quantities in (1) to (3)

$$x = \begin{bmatrix} x_0 \\ \zeta \end{bmatrix}; \; A = \begin{bmatrix} A_0 & B_0 \\ 0 & 0 \end{bmatrix}; \; B_1 = \begin{bmatrix} B_0 \\ 0 \end{bmatrix}; \; B_2 = \begin{bmatrix} B_0 \\ 0 \end{bmatrix} \tag{52}$$

$$C = \begin{bmatrix} C_0, & 0 \end{bmatrix} \tag{53}$$

$$D = \begin{bmatrix} D_0, & 1 \end{bmatrix} \tag{54}$$

where $(A_0,\ B_o)$ is assumed stabilizable, although, clearly, $(A,\ B_1)$ is not stabilizable

Our reason for choosing z as in (54) is that we wish to use z and $\hat{z}$ for the purpose of feedback control. Thus, we will assume that $(A_0 - B_0 D_0)$ is stable, i.e. the state feedback control law

$$u = -z = -Dx = -D_0 x_0 - \zeta \tag{55}$$

stabililzes the subset, x_0, of the states and cancels the input disturbance ζ leading to the closed loop response:

$$y = v \tag{56}$$

We next consider the estimation of z by a SEE of the form given in Figure 1 where $K(s)$ is partitioned as

$$K(s)^T = \left[K_1(s)^T,\ K_2(s)^T\right] \tag{57}$$

and where $B_1 u$ is injected into the open port on the left hand side of Figure 1.

Some simple algebra can be used to show that $D_0\hat{x}$ and $\hat{\zeta}$ have the following structure

$$D_0\hat{x} = \frac{s\gamma}{e}u + \frac{p}{e}y \tag{58}$$

$$\hat{\zeta} = \frac{-K_2 b}{e}u + \frac{K_2 a}{e}y \tag{59}$$

where

$$\frac{b}{a} = C_0(sI - A_0)^{-1}B_0 \tag{60}$$

$$\frac{s\gamma}{e} = D_0\left(sI - A_0 + K_1C_0 + B_0K_2/s\right)^{-1}B_0 \tag{61}$$

$$\frac{p}{e} = D_0\left(sI - A_0 + K_1C_0 + B_0K_2/s\right)^{-1}\left(K_1 + B_0K_2/s\right) \tag{62}$$

$$e = \left[\det\left(sI - A_0 + K_1C_0\right)\right]\left[s + K_2C\left(sI - A_0 + K_1C_0\right)^{-1}B_0\right] \tag{63}$$

where e has zeros inside the stability boundary by virtue of the SEE property.

Also, it is readily verified that the following unbiasedness constraint is automatically satisfied

$$\frac{d}{a} = \frac{s\gamma}{e} + \frac{pb}{ea} \tag{64}$$

where $\frac{d}{a} = D_0\left(sI - A_0\right)^{-1}B_0$.

The (certainty equivalence) state estimate feedback control law now takes the form:

$$u = -\hat{z} = -D_0\hat{x}_0 - \hat{\zeta} = -\hat{z}_0 - \hat{\zeta} \tag{65}$$

or

$$u = -\frac{s\gamma}{e}u - \frac{p}{e}y + \left[\frac{K_2 b}{e}u - \frac{K_2 a}{e}y\right] \tag{66}$$

This equation can be rearranged into the form:

$$\left[1 + \frac{s\gamma}{e} - \frac{K_2 b}{e}\right] u = -\left[\frac{p}{e} + \frac{K_2 a}{e}\right] y \tag{67}$$

It is readily checked that $(e - k_2 b)$ is zero for $s = 0$. Hence (67) takes the form:

$$\left[\frac{s\bar{\ell}}{e}\right] u = -\left[\frac{\bar{p}}{e}\right] y \tag{68}$$

where

$$s\bar{\ell} = [e + s\gamma - K_2 b] \tag{69}$$

$$\bar{p} = [p + K_2 a] \tag{70}$$

Actually, (64) can be rearranged into the form

$$ef = s\bar{\ell}a + \bar{p}b \tag{71}$$

where $f = (a + d)$ is stable by choice of D_0.

It then follows that the closed loop response is given by

$$y = \frac{b}{a+d}[\tilde{z} + w]$$

or

$$y = \frac{b}{a+d}\left\{[T_{\tilde{z}\omega} + 1]\, w - T_{\tilde{z}v} v\right\} \tag{72}$$

$$\begin{aligned} &= \frac{b}{a+d}\left\{\left[T_{\tilde{z}_0\omega} + T_{\tilde{\zeta}\omega} + 1\right] w - \left[T_{\tilde{z}_0 v} + T_{\tilde{\zeta}v}\right] v\right\} \\ &= \frac{b}{a+d}\left\{\left[\frac{s\gamma}{e} - \frac{K_2 b}{e} + 1\right] w + \left[\frac{p}{e} + \frac{K_2 a}{e}\right] v\right\} \\ &= \frac{b}{a+d}\left\{\left(\frac{s\bar{\ell}}{e}\right) w + \left(\frac{\bar{p}}{e}\right) v\right\} \end{aligned} \tag{73}$$

We define

$$S_i = \left(\frac{b}{a+d}\right)\left(\frac{s\bar{\ell}}{e}\right);\ S = \left(\frac{a}{b}\right) S_i,\ T = \left(\frac{b}{a+d}\right)\left(\frac{\bar{p}}{e}\right) \tag{74}$$

and note that from the unbiasedness constraint (70) we have the following complementary constraint of feedback control:

$$S + T = 1 \tag{75}$$

Several comments follow from the above development:

(i) S_i is the input sensitivity function of control (recall w is an input disturbance) whilst S and T are the standard sensitivity and complementary sensitivity functions of feedback control.

(ii) Equation (72) links filtering sensitivities to feedback control sensitivities. In particular, we see that poor sensitivity in the estimation of z is necessarily reflected in poor sensitivity in the closed loop response. In particular, the sensitivity S and complementary sensitivity T of control are related to the sensitivity functions of filtering by:

$$S = \left(\frac{a}{a+d}\right)[(T_{\tilde{z}\omega} + 1)] \tag{76}$$

$$T = -\left(\frac{b}{a+d}\right)(T_{\tilde{z}v}) \tag{77}$$

(iii) It is straightforward to use S and T to produce time and frequence constraints that must apply in linear feedback control. These parallel the results derived earlier in this paper for filtering. Full details are given in [2], [6], [8].

(iv) In the control context, we see from (72) that a desirable value for $T_{\tilde{z}\omega}$ is -1 at frequencies where we wish to be insensitive to ω. Including a deterministic disturbance, ζ, in the signal model (52) automatically leads to the constraint $T_{\tilde{z}\omega} = -1$ at d.c. as can be seen in (73). This suggests that it is not necessary to model the deterministic disturbance explicitly provided we add the constraint $(T_{\tilde{z}\omega} + 1) = 0$ at those frequencies at which deterministic input disturbance should be cancelled in closed loop.

(v) Finally (76), (77) suggest an alternative normalization of the filtering sensitivities without reference to feedback control, although f need not be stable in the filtering case.

11 Conclusions

This paper has shown that there exist fundamental constraints which apply to all filtering problems irrespective of the particular design method used. Links to corresponding results in feedback control have also been briefly reviewed.

References

[1] H. Bode, *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand, New York, 1945.

[2] J. Freudenberg and D. Looze, "Right half plane poles and zeros and design tradeoffs in feedback systems", *IEEE Trans. on Automatic Control*, Vol.AC-30, pp.555-565, June 1985.

[3] B. Francis and G. Zames, "On H_∞-optimal sensitivity theory for SISO feedback systems", *IEEE Trans. on Automatic Control*, Vol.AC-29, No.1, pp.9-16, 1984.

[4] S. O'Young and B. Francis, "Sensitivity tradeoffs for multivariable plants", *IEEE Trans. on Automatic Control*, Vol.AC-30, pp.625-632, July 1985.

[5] R.H. Middleton, "Trade-offs in linear control systems design", *Automatica*, Vol.27, pp.281-292, March 1991.

[6] J. Chen, "Sensitivity integral relations and design trade-offs in linear multivariable feedback systems", *IEEE Trans. on Automatic Control*, Vol.40, pp.1700-1716, October 1995.

[7] G.I. Gómez and G.C. Goodwin, "Vectorial sensitivity constraints for linear multivariable systems", in *Proceedings of the 34th CDC*, (New Orleans, LO), 1995.

[8] M.M. Seron, J. Braslavsky and G.C. Goodwin, *Fundamental Limitations in Filtering and Control.* Springer Verlag, 1996.

[9] G.C. Goodwin, D.Q. Mayne and J. Shim, "Trade-offs in linear filter design", *Automatica*, Vol.31, pp.1367-1376, October 1995.

[10] G.C. Goodwin and M.M. Seron, "Fundamental design trade-offs in filtering prediction and smoothing", *IEEE Trans. on Automatic Control*, to appear 1997.

[11] R. Grainger, *Feedback Issues in Filtering*, Ph.D. Thesis in preparation, The University of Newcastle, NSW Australia.

[12] G.C. Goodwin, A.R. Woodyatt, R.H. Middleton and J. Shim, "Fundamental limitations due to $j\omega$-Axis zeros in SISO systems", *Technical Report EE9723*, Department of Electrical and Computer Engineering, The University of Newcastle, Australia, 1997.

[13] G.C. Goodwin and R.H. Middleton, "The class of all stable unbiased estimators", *Systems and Control Letters*, Vol.13, pp.161-163, 1989.

[14] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach*, Prentice Hall Inc., Englewood Cliffs, NJ, 1990.

[15] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Prentice Hall Inc., Englewood Cliffs, NJ, 1979.

[16] T. Kailath, An innovations approach to least-squares estimation. Part I: Linear filtering in additive white noise", *IEEE Trans. on Automatic Control*, Vol.AC-13, No.6, pp.647-655, 1968.

[17] J. Braslavsky, M.M. Seron, D.Q. Mayne and P. Kokotovic, "Limiting performance of optimal linear filters", *Technical Report*, University of California, Santa Barbara, 1997.

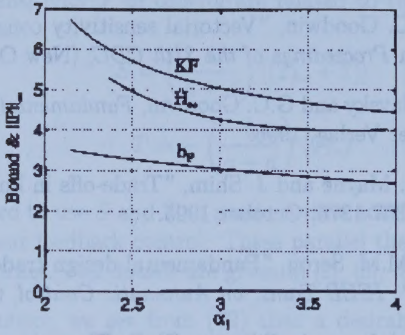


Figure 3: Achieved H_∞ norm for S_f based on KF and H_∞ designs, and predicted bound b_p.

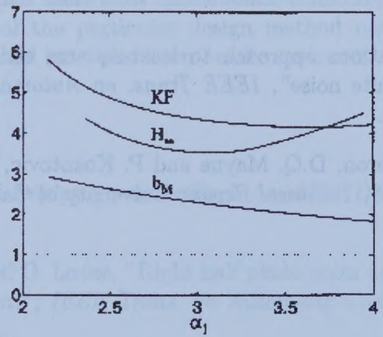


Figure 4: Achieved H_∞ norm for T_f based on KF and H_∞ designs, and predicted bound b_M.

DIFFERENTIAL TRAINING OF ROLLOUT POLICIES

by

Dimitri P. Bertsekas[2]

Abstract

We consider the approximate solution of stochastic optimal control problems using a neuro-dynamic programming/reinforcement learning methodology. We focus on the computation of a rollout policy, which is obtained by a single policy iteration starting from some known base policy and using some form of exact or approximate policy improvement. We indicate that, in a stochastic environment, the popular methods of computing rollout policies are particularly sensitive to simulation and approximation error, and we present more robust alternatives, which aim to estimate relative rather than absolute Q-factor and cost-to-go values. In particular, we propose a method, called *differential training*, that can be used to obtain an approximation to cost-to-go differences rather than cost-to-go values by using standard methods such as TD(λ) and λ-policy iteration. This method is suitable for recursively generating rollout policies in the context of simulation-based policy iteration methods.

1. INTRODUCTION

Consider the basic form of a dynamic programming problem, where we have the stationary discrete-time dynamic system

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots, \tag{1}$$

where x_k is the state taking values in some set, u_k is the control to be selected from a finite set $U(x_k)$, and w_k is a random disturbance. We assume that the disturbances w_k, $k = 0, 1, \ldots$, are independent, and take values in a countable set. The one-stage cost function is denoted by $g(x, u, w)$. Also for notational simplicity, we suppress time indexes for f, g, and U, but our discussion fully applies to a time-varying context where f, g, and U depend on the time index k.

We are given a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, which at time k maps a state x to a control $\mu_k(x) \in U(x)$. We focus primarily on an N-stage horizon problem where k takes the values $0, 1, \ldots, N-1$, and there is a terminal cost $G(x_N)$ that depends on the terminal state x_N, but our discussion

---

[1] Research supported by NSF under Grant 9625489-DMI. Thanks are due to John Tsitsiklis who suggested the analysis given at the end of Section 2.

[2] Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, Mass., 02139.

applies with minor modifications to various types of infinite horizon problems. The cost-to-go of π starting from an state x_k at time k will be denoted by

$$J_k(x_k) = E\left\{ G(x_N) + \sum_{i=k}^{N-1} g(x_i, \mu_i(x_i), w_i) \right\}. \tag{2}$$

The cost-to-go functions J_k satisfy the following recursion of dynamic programming (DP for short)

$$J_k(x) = E\big\{g(x, \mu_k(x), w) + J_{k+1}\big(f(x, \mu_k(x), w)\big)\big\}, \qquad k = 0, 1, \ldots \tag{3}$$

with the initial condition

$$J_N(x) = G(x).$$

Our discussion applies with minor modifications to infinite horizon problems. For such problems, we need to assume that the given policy is stationary of the form $\pi = \{\mu, \mu, \ldots\}$, in which case the DP algorithm is replaced by its asymptotic form (Bellman's equation). For example, if the problem is discounted with discount factor $\alpha \in (0, 1)$, the analog of Eq. (3) is

$$J(x) = E\big\{g(x, \mu(x), w) + \alpha J\big(f(x, \mu(x), w)\big)\big\}, \qquad \forall\ x,$$

where $J(x)$ is the cost-to-go of π starting from x.

The *rollout policy based on* π is denoted by $\overline{\pi} = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\}$, and is defined through the operation

$$\overline{\mu}_k(x) = \arg \min_{u \in U(x)} E\big\{g(x, u, w) + J_{k+1}\big(f(x, u, w)\big)\big\}, \qquad \forall\ x,\ k = 0, 1, \ldots. \tag{4}$$

Thus the rollout policy is a one step-lookahead policy, with the optimal cost-to-go approximated by the cost-to-go of the base policy. The name "rollout policy" was introduced by Tesauro [TeG96] in connection with one of his simulation-based computer backgammon algorithms. The book by Bertsekas and Tsitsiklis [BeT96] discusses in some detail various aspects of rollout policies in a stochastic context, and also in a deterministic combinatorial optimization context, as a device for magnifying the effectiveness of heuristics (see also Bertsekas, Tsitsiklis, and Wu [BTW97]).

A standard policy iteration argument can be used to show that the rollout policy $\overline{\pi}$ is an improved policy over the base policy π. In particular, let $\overline{J}_k(x)$ be the cost-to-go of the rollout policy $\overline{\pi}$ starting from a state x_k at time k. It can be shown that $\overline{J}_k(x) \leq J_k(x)$ for all x and k, so that the rollout policy $\overline{\pi}$ is an improved policy over the base policy π. The proof uses backwards induction on k. In particular, we have $\overline{J}_N(x) = J_N(x) = G(x)$ for all x. Assuming that $\overline{J}_{k+1}(x) \leq J_{k+1}(x)$ for all x, we have for all x,

$$\begin{aligned} \overline{J}_k(x) &= E\big\{g(x, \overline{\mu}_k(x), w) + \overline{J}_{k+1}\big(f(x, \overline{\mu}_k(x), w)\big)\big\} \\ &\leq E\big\{g(x, \overline{\mu}_k(x), w) + J_{k+1}\big(f(x, \overline{\mu}_k(x), w)\big)\big\} \\ &\leq E\big\{g(x, \mu_k(x), w) + J_{k+1}\big(f(x, \mu_k(x), w)\big)\big\} \\ &= J_k(x), \end{aligned}$$

where the first inequality follows from the induction hypothesis, the second inequality follows from Eq. (4), and the final equality follows from the DP equation (3). This completes the induction proof that $\overline{\pi}$ is an improved policy over π.

Rollout policies are useful in two main contexts:

(a) *The one-time policy improvement context.* Here π is some "easily" implementable heuristic policy. By this we mean that the controls $\mu_k(x)$ can be easily calculated for every state x through some known algorithm. Frequently the rollout policy is a considerable improvement over the base policy, so if π is already a fairly good heuristic policy, the rollout policy can be pretty close to being optimal and can form the basis for effective suboptimal control.

(b) *The policy iteration context.* Here the policy improvement operation is used multiple times to generate a sequence of policies, and aims at obtaining an optimal policy (or a close approximation thereof).

The one-time policy improvement approach is particularly useful in an *on-line control context*, where the problem is not fully known in advance, but rather it is only known to belong to a fairly broad problem class. The problem becomes fully specified to the controller just before the first control u_0 is to be chosen. In this case, it is difficult to use a policy iteration approach, but a one-time improvement using a rollout policy may be possible, as long as there is a base/heuristic policy that works for the entire problem class.

In practice, one typically has a method or algorithm to compute the control $\mu_k(x)$ of the base policy, given the state x, but the corresponding cost-to-go functions J_k may not be known in close form. Then the exact or approximate computation of the rollout control $\overline{\mu}_k(x)$ using Eq. (4) becomes an important and nontrivial issue, since we need for all $u \in U(x)$ the value of

$$Q_k(x,u) = E\{g(x,u,w) + J_{k+1}\big(f(x,u,w)\big)\}, \tag{5}$$

known as the *Q-factor* at time k. Alternatively, for the computation of $\overline{\mu}_k(x)$ we need the value of the cost-to-go $J_{k+1}\big(f(x,u,w)\big)$ at all possible next states $f(x,u,w)$.

We are interested in methods for approximately computing $\overline{\pi}$ in highly complex large-scale problems. Our assumption is that we can simulate the system under the base policy π, and in particular, that we can generate sample system trajectories and corresponding costs consistently with the probabilistic data of the problem. The standard methods of computing $\overline{\pi}$ involve the use of a function $\tilde{J}_k(x,r)$ approximating $J_k(x)$, where r is a tunable parameter vector that is obtained by various training methods. This is the Neuro-Dynamic Programming/Reinforcement Learning (NDP/RL for short) approach (see the survey by Barto, Bradtke, and Singh [BBS95], and the research monograph by Bertsekas and Tsitsiklis [BeT96], which list many references). A number of training methods are described in these references, including the TD(λ) method of Sutton [Sut88], and the Q-learning method of Watkins [Wat89]. There are rigorous convergence results for these methods, due to Tsitsiklis [Tsi94], and Tsitsiklis and Van Roy [TsV96], and described in [BeT96].

We argue, however, in this paper that the standard NDP/RL training techniques lack robustness to approximation and simulation error, and that this is a major difficulty in practical computation. Similar concerns have been voiced by Werbos [Wer92a], [Wer92b], who has proposed several algorithimc ideas in a variety of contexts, including continuous-time deterministic optimal control. The proposal of advantage updating by Baird, Harmon, and Klopf [Bai93], [HBK94], is similarly motivated by the sensitivity to simulation and approximation error in the computation of the rollout policy. The main idea of advantage updating is to compute an approximation $\tilde{A}_k(x, u, r)$ to the *advantage function*

$$A_k(x,u) = Q_k(x,u) - \min_{u \in U(x)} Q_k(x,u')$$

corresponding to state control pairs (x, u), where r is a tunable parameter vector. The (approximate) rollout control at state x is then computed by minimizing $\tilde{A}_k(x, u, r) \geq 0$ over $u \in U(x)$. One of the difficulties with advantage updating is that effective training methods such as TD(λ) cannot be used in a straightforward fashion to train the advantage function approximation $\tilde{A}_k(x, u, r)$.

In this paper, we discuss two methods for reducing the effects of the simulation and approximation error. Both of these methods may be viewed as variance reduction techniques adapted to the NDP/RL setting. In the first method, we consider the evaluation of Q-factor differences by Monte-Carlo simulation. In the second method, which we call *differential training*, we construct a function approximation $\tilde{G}_k(x, x', r)$ to the cost-to-go difference $J_k(x) - J_k(x')$, where x and x' is any pair of states, and r is a tunable parameter vector. An interesting aspect of differential training is that, contrary to advantage updating, it can use any of the standard NDP/RL methods, including TD(λ). This is accomplished by training on a special *differential system*, whose state is a pair of states (x, x') of the original system.

2. EVALUATING Q-FACTORS BY SIMULATION

A conceptually straightforward approach to compute the rollout control at a given state x and time k is to use Monte-Carlo simulation (this was Tesauro's original proposal in the context of backgammon [TeG96]). To implement this, we consider all possible controls $u \in U(x)$ and we generate a "large" number of simulated trajectories of the system starting from x, using u as the first control, and using the policy π thereafter. Thus a simulated trajectory has the form

$$x_{i+1} = f\big(x_i, \mu_i(x_i), w_i\big), \qquad i = k+1, \ldots, N-1.$$

where the first generated state is

$$x_{k+1} = f(x, u, w_k),$$

and each of the disturbances $w_k, \ldots, w_{N-1}$ is an independent random sample from the given distribution. The costs corresponding to these trajectories are averaged to compute an approximation $\tilde{Q}_k(x. u)$ to the Q-factor

$$Q_k(x, u) = E\big\{g(x, u, w) + J_{k+1}\big(f(x, u, w)\big)\big\}. \tag{6}$$

Note here that $\tilde{Q}_k(x, u)$ is an approximation to $Q_k(x, u)$ because of the simulation error resulting from the use of a limited number of trajectories. The approximation becomes increasingly accurate as the number of simulated trajectories increases. Once the approximate Q-factor $\tilde{Q}_k(x. u)$ corresponding to each control $u \in U(x)$ is computed, we can obtain the (approximate) rollout control $\tilde{\mu}_k(x)$ by the minimization

$$\tilde{\mu}_k(x) = \arg \min_{u \in U(x)} \tilde{Q}_k(x, u). \tag{7}$$

There is a serious flaw with this approach, due to the simulation error involved in the calculation of the Q-factors. In particular, for the calculation of $\tilde{\mu}_k(x)$ to be accurate, the Q-factor differences

$$Q_k(x, u) - Q_k(x, \hat{u})$$

must be computed accurately for all pairs of controls u and $\hat{u}$, so that these controls can be accurately compared. On the other hand, the simulation/approximation errors in the computation of the individual Q-factors $Q_k(x, u)$ are magnified through the preceding differencing operation.

An alternative idea is to approximate by simulation the Q-factor difference $Q_k(x. u) - Q_k(x, \hat{u})$ by sampling the difference

$$C_k(x, u, \mathbf{w}_k) - C_k(x, \hat{u}, \mathbf{w}_k),$$

where $\mathbf{w}_k = (w_k, w_{k+1}, \ldots, w_{N-1})$ and

$$C_k(x, u, \mathbf{w}_k) = G(x_N) + g(x, u, w_k) + \sum_{i=k+1}^{N-1} g\big(x_i, \mu_i(x_i), w_i\big). \tag{8}$$

This approximation may be far more accurate than the one obtained by differencing independent samples of $C_k(x, u, \mathbf{w}_k)$ and $C_k(x, \hat{u}, \mathbf{w}_k)$. Indeed, by introducing the zero mean sample errors

$$D_k(x, u, \mathbf{w}_k) = C_k(x, u, \mathbf{w}_k) - Q_k(x, u),$$

it can be seen that the variance of the error in estimating $Q_k(x,u) - Q_k(x,\hat{u})$ with the former method will be smaller than with the latter method if and only if

$$E_{\mathbf{w}_k,\,\hat{\mathbf{w}}_k}\left\{\left|D_k(x,u,\mathbf{w}_k) - D_k(x,\hat{u},\hat{\mathbf{w}}_k)\right|^2\right\} > E_{\mathbf{w}}\left\{\left|D_k(x,u,\mathbf{w}_k) - D_k(x,\hat{u},\mathbf{w}_k)\right|^2\right\}.$$

or equivalently

$$E\{D_k(x,u,\mathbf{w}_k)D_k(x,\hat{u},\mathbf{w}_k)\} > 0; \tag{9}$$

that is, if and only if there is positive correlation between the errors $D_k(x,u,\mathbf{w}_k)$ and $D_k(x,\hat{u},\mathbf{w}_k)$ (this idea was suggested by John Tsitsiklis). A little thought should convince the reader that this property is likely to hold in many types of problems. Roughly speaking, the relation (9) holds if changes in the value of u (at the first stage) have little effect on the value of the error $D_k(x,u,\mathbf{w}_k)$ relative to the effect induced by the randomness of $\mathbf{w}_k$. In particular, suppose that there exists a scalar $\gamma < 1$ such that, for all x, u, and $\hat{u}$, there holds

$$E\left\{\left|D_k(x,u,\mathbf{w}_k) - D_k(x,\hat{u},\mathbf{w}_k)\right|^2\right\} \le \gamma E\left\{\left|D_k(x,u,\mathbf{w}_k)\right|^2\right\}. \tag{10}$$

Then we have

$$\begin{aligned} D_k(x,u,\mathbf{w}_k)D_k(x,\hat{u},\mathbf{w}_k) &= \left|D_k(x,u,\mathbf{w}_k)\right|^2 + D_k(x,u,\mathbf{w}_k)\big(D_k(x,\hat{u},\mathbf{w}_k) - D_k(x,u,\mathbf{w}_k)\big) \\ &\ge \left|D_k(x,u,\mathbf{w}_k)\right|^2 - \left|D_k(x,u,\mathbf{w}_k)\right|\left|D_k(x,\hat{u},\mathbf{w}_k) - D_k(x,u,\mathbf{w}_k)\right|, \end{aligned}$$

from which we obtain

$$\begin{aligned} E\{D_k(x,u,\mathbf{w}_k)D_k(x,\hat{u},\mathbf{w}_k)\} &\ge E\left\{\left|D_k(x,u,\mathbf{w}_k)\right|^2\right\} \\ &\quad - E\left\{\left|D_k(x,u,\mathbf{w}_k)\right|\left|D_k(x,\hat{u},\mathbf{w}_k) - D_k(x,u,\mathbf{w}_k)\right|\right\} \\ &\ge E\left\{\left|D_k(x,u,\mathbf{w}_k)\right|^2\right\} \\ &\quad - \frac{1}{2}E\left\{\left|D_k(x,u,\mathbf{w}_k)\right|^2\right\} - \frac{1}{2}E\left\{\left|D_k(x,\hat{u},\mathbf{w}_k) - D_k(x,u,\mathbf{w}_k)\right|^2\right\} \\ &\ge \frac{1-\gamma}{2}E\left\{\left|D_k(x,u,\mathbf{w}_k)\right|^2\right\}. \end{aligned}$$

Thus, under the assumption (10) and the assumption $E\left\{\left|D_k(x,u,\mathbf{w}_k)\right|^2\right\} > 0$, the condition (9) holds and guarantees that by directly obtaining cost difference samples rather than differencing cost samples, the error variance decreases.

3. APPROXIMATING COST-TO-GO DIFFERENCES

A common approach to compute an approximate rollout policy $\overline{\pi}$ consists of two steps:

(1) In a preliminary phase, approximations to the cost-to-go functions J_k, denoted $\tilde{J}_k$, are computed, possibly using simulation and a least squares fit from a parametrized class of functions.

The techniques of NDP/RL such as temporal difference methods and others can be used for the training of such an approximation (see [BeT96] for a detailed account).

(2) Given $\tilde{J}_k$ and a state x at time k, the approximate Q-factor

$$\tilde{Q}_k(x,u) = E\{g(x,u,w) + \tilde{J}_{k+1}(f(x,u,w))\}$$

is computed for all $u \in U(x)$, and the (approximate) rollout control $\overline{\mu}_k(x)$ is obtained by the minimization

$$\overline{\mu}_k(x) = \arg \min_{u \in U(x)} \tilde{Q}_k(x,u).$$

Unfortunately, this method also suffers from the error magnification inherent in the Q-factor differencing operation. This motivates an alternative approach, called *differential training*, that is based on cost-to-go difference approximation. We observe that to compute the rollout control $\overline{\mu}_k(x)$, it is sufficient to have the differences of costs-to-go

$$J_{k+1}(f(x,u,w)) - J_{k+1}(f(x,\hat{u},w)).$$

Thus, we consider approximating this difference using a function approximator, which given any two states s and $\hat{s}$, gives an approximation $\tilde{G}_{k+1}(s,\hat{s})$ of $J_{k+1}(s) - J_{k+1}(\hat{s})$. The rollout control can then be approximated by

$$\overline{\mu}_k(x) = \arg \min_{u \in U(x)} E\{g(x,u,w) - g(x,\mu(x),w) + \alpha\tilde{G}_{k+1}(f(x,u,w), f(x,\mu_k(x),w))\}. \tag{11}$$

An important point here is that the training of an approximation architecture to obtain $\tilde{G}_{k+1}$ can be done using any of the standard NDP/RL methods [e.g., TD(λ), approximate and optimistic policy iteration or λ-policy iteration, etc.]. To see this, denote for all k

$$G_k(x,\hat{x}) = J_k(x) - J_k(\hat{x}).$$

By subtracting the DP equations corresponding to π, and to x and $\hat{x}$,

$$J_k(x) = E\{g(x,\mu_k(x),w) + J_{k+1}(f(x,\mu_k(x),w))\},$$

$$J_k(\hat{x}) = E\{g(\hat{x},\mu_k(\hat{x}),w) + J_{k+1}(f(\hat{x},\mu_k(\hat{x}),w))\},$$

we obtain, for all $(x,\hat{x})$ and k,

$$G_k(x,\hat{x}) = E\{g(x,\mu_k(x),w) - g(\hat{x},\mu_k(\hat{x}),w) + G_{k+1}(f(x,\mu_k(x),w), f(\hat{x},\mu_k(\hat{x}),w))\}.$$

Therefore, G_k can be viewed as the cost-to-go function for a problem involving a fixed policy, the state $(x, \hat{x})$, the cost per stage

$$g(x, \mu_k(x), w) - g(\hat{x}, \mu_k(\hat{x}), w), \tag{12}$$

and the state transition equation

$$(x_{k+1}, \hat{x}_{k+1}) = \big(f(x_k, \mu_k(x_k), w_k), f(\hat{x}_k, \mu_k(\hat{x}_k), w_k)\big). \tag{13}$$

Thus, it can be seen that *any of the standard methods that can be used to train architectures that approximate J_k, can also be used for training architectures that approximate G_k.* For example, one may use simulation-based methods, such as TD(λ), that generate trajectories consisting of pairs $(x, \hat{x})$ according to Eq. (13). Note that a single random sequence $\{w_0, \ldots, w_{N-1}\}$ may be used to simultaneously generate samples of $G_k(x, \hat{x})$ for several triples $(x, \hat{x}, k)$. This is similar in spirit to the ideas of Perturbation Analysis (see e.g., Cassandras [Cas93], Glasserman [Gla91], or Ho and Cao [HoC91]).

The approach just described is well suited to a policy iteration context, where subsequent to the relative cost-to-go approximation corresponding to the base policy π, we will be interested in constructing a relative cost-to-go approximation corresponding to the rollout policy $\overline{\pi}$. For this construction, simulation will be used, during which it is essential that the next policy $\overline{\pi}$ can be fairly easily generated. Using Eq. (11) is quite convenient for this purpose.

A special case of interest arises when a linear, feature-based architecture is used for the approximator $\tilde{G}_k$. In particular, let F_k be a feature extraction mapping that associates a feature vector $F_k(s)$ with state s and time k, and let $\tilde{G}_k$ be of the form

$$\tilde{G}_k(s, \hat{s}) = r(k)'\big(F_k(s) - F_k(\hat{s})\big),$$

where $r(k)$ is a tunable weight vector of the same dimension as $F_k(s)$ and prime denotes transposition. The rollout policy is generated by

$$\overline{\mu}_k(x) = \arg \min_{u \in U(x)} E\big\{g(x, u, w) + r(k)' F_k\big(f(x, u, w)\big)\big\},$$

which corresponds to using $r(k)'F_k(s)$ (plus an unknown inconsequential constant) as an approximator to $J_k(s)$. Thus, in this approach, *we essentially use a linear, feature-based architecture to approximate J_k, but we train this architecture using the differential system (13) and the differential cost per stage (12).*

The preceding discussion also applies to infinite horizon problems. The analog of Eq. (11) is

$$\overline{\mu}(x) = \arg \min_{u \in U(x)} E\big\{g(x, u, w) - g(x, \mu(x), w) + \alpha \tilde{G}\big(f(x, u, w), f(x, \mu(x), w)\big)\big\}.$$

and $\tilde{G}$ is an approximation to the cost-to-go difference, defined for any pair of states $(x, \hat{x})$ by

$$G(x, \hat{x}) = J(x) - J(\hat{x}).$$

By subtracting the two Bellman's equations corresponding to μ, and to x and $\hat{x}$,

$$J(x) = E\{g(x, \mu(x), w) + \alpha J(f(x, \mu(x), w))\},$$

$$J(\hat{x}) = E\{g(\hat{x}, \mu(\hat{x}), w) + \alpha J(f(\hat{x}, \mu(\hat{x}), w))\},$$

we obtain

$$G(x, \hat{x}) = E\{g(x, \mu(x), w) - g(\hat{x}, \mu(\hat{x}), w) + \alpha G(f(x, \mu(x), w), f(\hat{x}, \mu(\hat{x}), w))\}, \quad \forall\ (x, \hat{x}).$$

Therefore, G can be viewed as the cost-to-go function for a problem involving a fixed policy, the state $(x, \hat{x})$, the cost per stage

$$g(x, \mu(x), w) - g(\hat{x}, \mu(\hat{x}), w), \tag{14}$$

and the state transition equation

$$(x_{k+1}, \hat{x}_{k+1}) = (f(x_k, \mu(x_k), w_k), f(\hat{x}_k, \mu(\hat{x}_k), w_k)). \tag{15}$$

The simulation-based training of $\tilde{G}$ can be done by using the differential cost (14) and the differential system (15).

4. CONCLUSIONS

The main theme of this paper, namely that one should try to approximate Q-factor and cost-to-go differences rather than Q-factors has been emphasized by a number of authors. In particular, it has been discussed by Werbos [Wer92a], [Wer92b] in several contexts, including continuous-time deterministic optimal control, and by Baird, Harmon, and Klopf [Bai93], [HBK94] in the context of advantage updating. The differential training method proposed in this paper, is new to the author's knowledge, and has the significant advantage that it requires straightforward adaptations of well-known and tested methods such as TD(λ) and λ-policy iteration. We only have preliminary computational experience with the differential training method, which is however, impressively positive, and has been successful in problems where the standard training methods have failed. A somewhat restrictive assumption in our discussion is the independence of the disturbances w_k. In many problems of interest, w_k depends on preceding disturbances via the current state x_k and/or the control u_k. On the other hand, one can often introduce approximations in the approximation model to eliminate such dependencies.

REFERENCES

[BBS95] Barto, A. G., Bradtke, S. J., and Singh, S. P., 1995. "Learning to Act Using Real-Time Dynamic Programming," Artificial Intelligence, Vol. 72, pp. 81-138.

[BTW97] Bertsekas, D. P., Tsitsiklis, J. N., and Wu, C., 1997. "Rollout Algorithms for Combinatorial Optimization," Report LIDS-P-2386, Lab. for Information and Decision Systems, Mass. Institute of Technology, Cambridge, MA; to appear in Heuristics.

[Bai93] Baird, L. C., 1993. "Advantage Updating," Report WL-TR-93-1146, Wright Patterson AFB, OH.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA.

[Cas93] Cassandras, C. G., 1993. "Discrete-Event Systems: Modeling and Performance Analysis," Aksen Associates, Boston, MA.

[Gla91] Glasserman, P., 1991. "Gradient Estimation via Perturbation Analysis," Kluwer, Boston, MA.

[HBK94] Harmon, M. E., Baird, L. C., and Klopf, A. H., 1994. "Advantage Updating Applied to a Differential Game," unpublished report, presented at the 1994 Neural Information Processing Systems Conference, Denver, CO.

[HoC91] Ho, Y. C., and Cao, X., 1991. "Perturbation Analysis of Discrete-Event Systems," Kluwer, Boston, MA.

[TeG96] Tesauro, G., and Galperin, G. R., 1996. "On-Line Policy Improvement Using Monte Carlo Search," unpublished report, presented at the 1996 Neural Information Processing Systems Conference, Denver, CO.

[TsV96] Tsitsiklis, J. N., and Van Roy, B., 1996. "An Analysis of Temporal-Difference Learning with Function Approximation," Lab. for Info. and Decision Systems Report LIDS-P-2322, Massachusetts Institute of Technology, Cambridge, MA.

[Tsi94] Tsitsiklis, J. N., 1994. "Asynchronous Stochastic Approximation and Q-Learning," Machine Learning, Vol. 16, pp. 185-202.

[Wer92a] Werbös, P. J, 1992. "Approximate Dynamic Programming for Real-Time Control and Neural Modeling," D. A. White, D. A. Sofge, (eds.), Handbook of Intelligent Control, Van Nostrand, N. Y.

[Wer92b] Werbös, P. J, 1992. "Neurocontrol and Supervised Learning: an Overview and Valuation," D. A. White, D. A. Sofge, (eds.), Handbook of Intelligent Control, Van Nostrand, N. Y.

ESTIMATION OF STOCHASTIC-PARAMETER SIGNALS USING LINEAR MATRIX INEQUALITIES[1]

Jafar Mohseni, Edwin E. Yaz[2], Kraig J. Olejniczak
Department of Electrical Engineering
University of Arkansas
Fayetteville, AR 72701-1201

Abstract- The objective of this paper is to propose the use of linear matrix inequalities for estimation of random-parameter signals modeled in state space with finite energy-type disturbances. The signal to be estimated is assumed to have random parameters and is corrupted by finite-energy measurement and process disturbances which are either deterministic or stochastic in nature. Estimation techniques based on two criteria are presented to predict the signal when corrupted by these disturbances. Two examples are provided to illustrate the proposed techniques.

I. INTRODUCTION

In this work, we present linear matrix inequality (LMI)-based techniques to find the filter gains in estimating a signal with a random-parameter model which is corrupted by finite-energy process and measurement disturbances. The filter gain, which is constant, can be used in on-line estimators. This is a novel approach to obtaining the gain of state estimators for stochastic-parameter signals.

Stochastic-parameter systems have been used in different applications. Systems with white parameters have been used, for example, in satellite attitude control [1], chemical reactor control [2], macroeconomics [3], and in robustification of controllers [4]. Other applications include modeling the intermittent measurement problem where the signal content of the measurement may be absent at times, such as in radar tracking [5], and in random-amplitude modulation problems.

Other works on state estimation of discrete-time systems with white parameters include [6]-[9]. The conventional approach to this problem is to perceive it from the minimum variance optimality viewpoint [6]. However, [7] is on suboptimal observer design with guaranteed stochastic convergence. Robustness results are presented in [8] and [9] contains a fixed-order observer design procedure for this class of systems. Applicability of these observers to more general stochastic models is shown in [10].

In this paper, we will introduce a model for random-parameter signals using deterministic or stochastic finite-energy ($\mathcal{L}_2$-type) disturbance signals and then apply and assess the performance of LMI-based estimators. The use of $\mathcal{L}_2$-type signals in modeling is motivated by the realistic assumption that the ambient noise or disturbance effects have finite energy and the choice of LMI-based estimators is dictated by the availability of efficient numerical methods for their solution. State estimators with guaranteed estimation error bounds, and with bounds on the energy of the estimation error vector, will be introduced. An example will be provided to illustrate the performance of these estimators.

---

[1] This research was partially supported by the NSF under the grant number ECS-9322798.

[2] The corresponding author.

The following notation is used. For an n-dimensional vector of real variables $x \in \mathbb{R}^n$, $\|x\|$ denotes the 2-norm, $(x^T x)^{1/2}$. For an $n \times n$ symmetric matrix A with real elements $A \in \mathbb{R}^{n \times n}$, $A > 0$ ($A < 0$), $\lambda_{max}(A)$ ($\lambda_{min}(A)$) respectively mean positive definite (negative definite) A and its maximum (minimum) eigenvalue. $\mathscr{E}\{x\}$ and $\overline{x}$ alternatively denote the expected value of x.

II. ESTIMATION

Before introducing our estimators, a preliminary result that will help us in developing our design equations will be given below:

Lemma 1. Consider the state-space model

$$X_{k+1} = \Phi_k X_k + \Gamma_k w_k$$
$$Z_k = \Psi_k X_k + \Lambda_k w_k \tag{1}$$

where all the model parameters Φ_k, Γ_k, Ψ_k, and Λ_k are assumed to be weakly stationary, white (uncorrelated in time) and uncorrelated with X_0 and w_k is a deterministic unit energy signal ($\sum_{k=0}^{\infty} \|w_k\|^2 \leq 1$). If there exists a $P > 0$ and $\gamma > 0$ such that

$$\begin{bmatrix} P & 0 \\ 0 & \gamma I \end{bmatrix} - \mathscr{E}\left\{ \begin{bmatrix} \Phi^T \\ \Gamma^T \end{bmatrix} P \begin{bmatrix} \Phi & \Gamma \end{bmatrix} \right\} > 0 \tag{2}$$

then the state X_k satisfies

$$\mathscr{E}\{X_N^T P X_N\} < \gamma + \mathscr{E}\{X_0^T P X_0\} \tag{3}$$

for any $N \geq 0$. Also, if there exists a $P > 0$ and $\gamma > 0$ such that

$$\begin{bmatrix} P & 0 \\ 0 & \gamma I \end{bmatrix} - \mathscr{E}\left\{ \begin{bmatrix} \Phi^T \\ \Gamma^T \end{bmatrix} P \begin{bmatrix} \Phi & \Gamma \end{bmatrix} - \begin{bmatrix} \Psi^T \\ \Lambda^T \end{bmatrix} \begin{bmatrix} \Psi & \Lambda \end{bmatrix} \right\} > 0 \tag{4}$$

then for $N \geq 0$, the output energy of this model satisfies

$$\sum_{k=0}^{N} \mathscr{E}\{\|Z_k\|^2\} < \gamma + \mathscr{E}\{X_0^T P X_0\} \tag{5}$$

Proof. Define $V_k = X_k^T P X_k$ for some $P > 0$ and consider

$$\{V_{k+1} - V_k - \gamma \|w_k\|^2\} = \mathscr{E}\left\{ \begin{bmatrix} X_k^T & w_k^T \end{bmatrix} \begin{bmatrix} \Phi_k^T P \Phi_k - P & \Phi_k^T P \Gamma_k \\ \Gamma_k^T P \Phi_k & \Gamma_k^T P \Gamma_k - \gamma I \end{bmatrix} \begin{bmatrix} X_k \\ w_k \end{bmatrix} \right\} \tag{6}$$

$$= \mathscr{E}\left\{ \begin{bmatrix} X_k^T & w_k^T \end{bmatrix} \right\} \mathscr{E}\left\{ \begin{bmatrix} \Phi^T P \Phi - P & \Phi^T P \Gamma \\ \Gamma^T P \Phi & \Gamma^T P \Gamma - \gamma I \end{bmatrix} \right\} \mathscr{E}\left\{ \begin{bmatrix} X_k \\ w_k \end{bmatrix} \right\}$$

that easily follows by substituting from (1) and using the property that the state vector is uncorrelated with the parameters of the model. In this case, LMI (2) yields

$$\mathscr{E}\{\ V_{k+1} - V_k\ \} - \gamma \|w_k\|^2 < 0 \tag{7}$$

By summing over k we get

$$\sum_{k=0}^{N-1} \mathscr{E}\{\ V_{k+1} - V_k\ \} = \mathscr{E}\{\ V_N - V_0\ \} < \gamma \sum_{k=0}^{N-1} \|w_k\|^2 \le \gamma \tag{8}$$

that gives the result (3). Also, considering

$$\mathscr{E}\left\{V_{k+1} - V_k + \|Z_k\|^2 - \gamma \|w_k\|^2\right\}$$

$$= \mathscr{E}\left\{[X_k^T \quad w_k^T]\right\} \mathscr{E}\left\{\begin{bmatrix} \Phi^T P \Phi - P + \Psi^T \Psi & \Phi^T P \Gamma + \Psi^T \Lambda \\ \Gamma^T P \Phi + \Lambda^T \Psi & \Gamma^T P \Gamma - \gamma I + \Lambda^T \Lambda \end{bmatrix}\right\} \mathscr{E}\left\{\begin{bmatrix} X_k \\ w_k \end{bmatrix}\right\} \tag{9}$$

which can be easily verified. By assuming that the random parameters of the model are uncorrelated with the state and disturbance vectors, LMI (4) yields

$$\mathscr{E}\{\ V_{k+1} - V_k + \|Z_k\|^2 - \gamma \|w_k\|^2\ \} < 0 \tag{10}$$

By summing over k we get

$$\sum_{k=0}^{N} \mathscr{E}\{\ V_{k+1} - V_k + \|Z_k\|^2\ \} < \gamma \sum_{k=0}^{N} \|w_k\|^2 \tag{11}$$

or

$$\sum_{k=0}^{N} \mathscr{E}\{\ \|Z_k\|^2\ \} < \gamma + V_0 \tag{12}$$

which gives the result (5).

Lemma 2. Consider the model (1) with the same description for parameters but w_k is assumed to be a stochastic unit-energy signal, ($\sum_{k=0}^{\infty} \mathscr{E}\{\|w_k\|^2\} \le 1$) which is uncorrelated with both the parameters of the model and X_0. Then the same conditions (2) and (4) will imply the respective results (3) and (5).

Proof. The proof follows the same lines as in Lemma 1, except that (7) is replaced by

$$\mathscr{E}\{\ V_{k+1} - V_k - \gamma \|w_k\|^2\ \} < 0$$

and (11) is replaced by

$$\sum_{k=0}^{N} \mathscr{E}\{\ V_{k+1} - V_k + \|Z_k\|^2 - \gamma \sum_{k=0}^{N} \|w_k\|^2\ \} < 0$$

Let us now apply these results to estimator design. Consider the following state-space representation for a random-parameter signal

$$\begin{aligned} x_{k+1} &= A_k\, x_k + F w_k \\ y_k &= C x_k + G w_k \end{aligned} \tag{13}$$

where A_k is a matrix of weakly stationary white noises uncorrelated with x_0 which has the mean $\bar{x}_0$ and w_k is a deterministic or stochastic unit-energy disturbance. We use a linear, unbiased and full-order filter to estimate the state vector x_k

$$\hat{x}_{k+1} = \bar{A}\hat{x}_k + K\,(\,y_k - C\hat{x}_k\,), \quad \hat{x}_0 = \bar{x}_0 \tag{14}$$

Forming the error equation, we arrive at

$$e_{k+1} = x_{k+1} - \hat{x}_{k+1} = (\bar{A} - KC)e_k + \tilde{A}_k x_k + (F - KG)w_k, \quad \tilde{A}_k \doteq A_k - \bar{A} \tag{15}$$

Combining (13) and (15), we get the following state-space representation

$$X_k = \begin{bmatrix} x_{k+1} \\ e_{k+1} \end{bmatrix} = \begin{bmatrix} A_k & 0 \\ \tilde{A}_k & \bar{A} - KC \end{bmatrix} \begin{bmatrix} x_k \\ e_k \end{bmatrix} + \begin{bmatrix} F \\ F - KG \end{bmatrix} w_k \tag{16}$$

and since the behavior of error e_k is of interest, we can define

$$Z_k = e_k = T \begin{bmatrix} x_k \\ e_k \end{bmatrix}, \quad T = [\ 0 \quad I\ ] \tag{17}$$

Comparing (16) and (17) with (1), we see for this model we have

$$\Phi_k = \begin{bmatrix} A_k & 0 \\ \tilde{A}_k & \bar{A} - KC \end{bmatrix}, \quad \Gamma_k = \Gamma = \begin{bmatrix} F \\ F - KG \end{bmatrix} \tag{18}$$
$$\Psi_k = \Psi = [\ 0 \quad I\ ], \quad \Lambda_k = \Lambda = 0$$

In order to find the estimator gain using the above-mentioned LMIs, we define $V_k = x_k^T P_1 x_k + e_k^T P_2 e_k$ with $P_1 > 0$ and $P_2 > 0$, which means that the P matrix is block diagonal with P_1 and P_2 as its diagonal blocks. We use the above values for the matrices Φ_k, Γ_k, Ψ_k, and Λ_k in LMI (2). The following equations can easily be verified:

$$\Phi_k^T P \Phi_k = \begin{bmatrix} A_k^T P_1 A_k + \tilde{A}_k^T P_2 \tilde{A}_k & \tilde{A}_k^T P_2 (\bar{A} - KC) \\ (\bar{A} - KC)^T P_2 \tilde{A}_k & (\bar{A} - KC)^T P_2 (\bar{A} - KC) \end{bmatrix} \tag{19}$$

$$\Phi_k^T P \Gamma_k = \begin{bmatrix} A_k^T P_1 F + \tilde{A}_k^T P_2 (F - KG) \\ (A - KC)^T P_2 (F - KG) \end{bmatrix} \tag{20}$$

$$\Gamma_k^T P \Gamma_k = F^T P_1 F + (F - KG)^T P_2 (F - KG) \tag{21}$$

Substituting in LMI (2) and taking the expected value of each block in the matrix yield

$$\begin{bmatrix} P_1 - \overline{A^T P_1 A} - \overline{\tilde{A}^T P_2 \tilde{A}} & 0 & -\bar{A}^T P_1 F \\ 0 & P_2 & 0 \\ -F^T P_1 \bar{A} & 0 & \gamma I - F^T P_1 F \end{bmatrix} - \begin{bmatrix} 0 \\ (\bar{A} - KC)^T \\ (F - KG)^T \end{bmatrix} P_2 \begin{bmatrix} 0 & \bar{A} - KC & F - KG \end{bmatrix} > 0 \tag{22}$$

Since for any G_{11}, G_{12}, and G_{22} having

$$\begin{bmatrix} G_{11} & G_{12} \\ G_{12}^T & G_{22} \end{bmatrix} > 0$$

is equivalent to $G_{22} > 0$ and $G_{11} - G_{12} G_{22}^{-1} G_{12}^T > 0$, (22) is the same as

$$\begin{bmatrix} P_1 - \overline{A^T P_1 A} - \overline{\tilde{A}^T P_2 \tilde{A}} & 0 & -\bar{A}^T P_1 F & 0 \\ 0 & P_2 & 0 & (\bar{A} - KC)^T P_2 \\ -F^T P_1 \bar{A} & 0 & \gamma I - F^T P_1 F & (F - KG)^T P_2 \\ 0 & P_2 (\bar{A} - KC) & P_2 (F - KG) & P_2 \end{bmatrix} > 0 \tag{23}$$

which can be written with $Y = P_2 K$ as

$$\begin{bmatrix} P_1 - \overline{A^T P_1 A} - \overline{\tilde{A}^T P_2 \tilde{A}} & 0 & -\bar{A}^T P_1 F & 0 \\ 0 & P_2 & 0 & \bar{A}^T P_2 - C^T Y^T \\ -F^T P_1 \bar{A} & 0 & \gamma I - F^T P_1 F & F^T P_2 - G^T Y^T \\ 0 & P_2 \bar{A} - YC & P_2 F - YG & P_2 \end{bmatrix} > 0 \qquad (24)$$

LMI (24) can be solved for $P_1 > 0$, $P_2 > 0$, $\gamma > 0$, and Y. Once these values are calculated, then the estimator gain K can be found by $K = P_2^{-1} Y$.

To find the equivalent of LMI (4) for the system described by (13), we adopt the same approach as elucidated for LMI (2) and obtain the following LMI

$$\begin{bmatrix} P_1 - \overline{A^T P_1 A} - \overline{\tilde{A}^T P_2 \tilde{A}} & 0 & -\bar{A}^T P_1 F & 0 \\ 0 & P_2 - I & 0 & \bar{A}^T P_2 - C^T Y^T \\ -F^T P_1 \bar{A} & 0 & \gamma I - F^T P_1 F & F^T P_2 - G^T Y^T \\ 0 & P_2 \bar{A} - YC & P_2 F - YG & P_2 \end{bmatrix} > 0 \qquad (25)$$

Note that the only difference between (24) and (25) is in the block located in the second row and column. Similar to (24), in LMI (25) also we solve for $P_1 > 0$, $P_2 > 0$, $\gamma > 0$, and Y and then find the estimator gain K from $K = P_2^{-1} Y$.

To summarize our development, consider the system described by (13). Then we have:

Theorem 1. Let LMI (24) hold for $P_1 > 0$, $P_2 > 0$, $\gamma > 0$, and Y. Then, there exists a full-order estimator (14) such that the vector composed of the state and the estimation error satisfies (3).

Let LMI (25) hold for $P_1 > 0$, $P_2 > 0$, $\gamma > 0$, and Y. Then, there exists a full-order estimator (14) such that the estimation error satisfies (5).

For the system (13), inequalities (3) and (5), respectively, yield the following bounds

$$\begin{aligned} &\lambda_{\min}(P_1)\mathscr{E}\{\|x_N\|^2\} + \lambda_{\min}(P_2)\mathscr{E}\{\|e_N\|^2\} \\ &\leq \mathscr{E}\{\, x_N^T P_1 x_N \,\} + \mathscr{E}\{\, e_N^T P_2 e_N \,\} \\ &\leq \gamma + \mathscr{E}\{\, x_0^T P_1 x_0 \,\} + \mathscr{E}\{\, e_0^T P_2 e_0 \,\} \end{aligned} \qquad (26)$$

$$\sum_{k=0}^{N} \mathscr{E}\{ \|e_k\|^2 \} < \gamma + \lambda_{\max}(P_1)\, \mathscr{E}\{ \|x_0\|^2 \} + \lambda_{\max}(P_2)\, \mathscr{E}\{ \|e_0\|^2 \} \tag{27}$$

Note that matrix P in inequalities (3) and (5) is assumed to be a block diagonal matrix with P_1 and P_2 as its diagonal blocks. Inequality (26) suggests that we maximize $\lambda_{min}(P_1)$ and $\lambda_{min}(P_2)$ and minimize γ. Likewise, (27) suggests that we minimize $\lambda_{max}(P_1)$ and $\lambda_{max}(P_2)$ and γ. These optimization criteria together with LMIs (26) and (27) will result in eigenvalue problems [11].

III. EXAMPLE AND SIMULATION RESULTS

To demonstrate the performance of the proposed estimators, we present a signal which can be modeled using a random-parameter state-space representation. After introducing the model for the signal, the simulation results in estimating this signal will be shown.

The example signal is a sinusoid whose frequency is a binary random variable and whose amplitude is decaying slowly, along with some measurement noise.

$$s_k = A_k \sin(\omega_k kT + \theta) + w_k \tag{28}$$

where $\omega_k = \omega_1$ with the probability p_1 and $\omega_k = \omega_2$ with the probability $p_2 = 1 - p_1$. w_k is the finite-energy measurement noise. With the choice of $x_k = [\, A_0 \sin(\omega_k kT + \theta),\, A_0 \cos(\omega_k kT + \theta)\,]^T$ it can easily be seen that the above signal can be modeled in the following state-space representation.

$$\begin{aligned} x_{k+1} &= \alpha \begin{bmatrix} \cos\omega_k T & \sin\omega_k T \\ -\sin\omega_k T & \cos\omega_k T \end{bmatrix} x_k \\ y_k &= [1 \quad 0]\, x_k + w_k \end{aligned} \tag{29}$$

where α is a constant which accounts for the decay in the amplitude of the sinusoid. This signal is depicted in Figure 1 for $\omega_1 = 100\pi$, $p_1 = 0.9$, $\omega_2 = 200\pi$, $p_2 = 0.1$, $T = 1/200$ sec, and $\alpha = 0.99$. We chose the initial state x_0 randomly from a uniform distribution on [0, 1]. To make the filter unbiased, the initial state estimate $\hat{x}_0$ should be the mean value of x_0. The noise w_k is generated by putting an energy bound on the output of a pseudo-random Gaussian-distributed number generator.

To build the estimators, LMIs (24) and (25) were solved for P_1, P_2, γ, and Y under their corresponding optimization criteria, given by inequalities (26) and (27), respectively. It is worth mentioning that since it was not possible with the available software to maximize the minimum eigenvalue of a matrix or minimize its maximum eigenvalue directly, other equivalent criteria were used. Maximizing the minimum eigenvalue of a matrix S is equivalent to maximizing δ such that $S - \delta I > 0$; and minimizing the maximum eigenvalue of S is equivalent to minimizing δ such that $\delta I - S > 0$ [11]. After finding the optimal values, the estimation gain K was simply obtained from $K = P_2^{-1} Y$. All of the following illustrated simulation results are the Monte Carlo average of 100 runs.

Figures 2 and 4 represent the error signals $\mathcal{E}\{y_k - \hat{y}_k\}$ for both estimators. Also, the expected value of the error bound $\mathcal{E}\{x_k^T P_1 x_k + e_k^T P_2 e_k\}$ for the first estimator and the expected value of the error energy $\sum \mathcal{E}\{\|e_k\|^2\}$ for the second estimator are shown in Figures 3 and 5 to illustrate that they are bounded, as expected by the inequalities (3) and (5).

It is informative to compare the performance of the proposed techniques with the conventional Kalman filter. Since in a Kalman filter the parameters of the model cannot be random, the measurement signal in Figure 1 is modeled using a representation similar to (29) but with only one frequency, i.e., the more likely frequency ω_1. Figure 6 shows $\sum \mathcal{E}\{\|e_k\|^2\}$ for the Kalman filter and comparing with Figure 5, it is evident that the Kalman filter renders a higher error energy. Our proposed estimators have the additional advantage of having constant gains, whereas in the Kalman filter case the gain needs to be updated at each step.

IV. CONCLUSION

Two estimator designs have been proposed for linear, discrete-time, random-parameter state-space models which are corrupted by finite energy-type deterministic or stochastic disturbances. These estimator equations have been derived based on keeping either a weighted quadratic function of the estimation error or the energy of the estimator error small. These two estimators have been applied to a problem involving one-step-ahead prediction of a sinusoid with a binary random frequency modeled as a random-parameter system. Simulation results validate the efficacy of the estimators.

REFERENCES

[1] P. J. McLane, "Optimal Stochastic Control of Linear Systems with State and Control Dependent Disturbances," *IEEE Trans. Automatic Control*, vol. 16, pp. 793-798, 1971.

[2] N. J. Rao, D. Ramakrishna, D. J. Borwanker, "Nonlinear Stochastic Simulation of Stirred Tank Reactors," *Chemical Engineering Science*, vol. 29, pp. 1193-1204, 1974.

[3] M. Aoki, *Optimal Control and System Theory in Dynamic Economic Analysis*, Elsevier: New York, 1976.

[4] E. Yaz, "Feedback Controllers for Stochastic-Parameter Systems: Relationships among Various Stabilizability Conditions," *Optimal Control Application Methods*, vol. 9, pp. 325-332, 1988.

[5] W. NaNacara and E. Yaz, "Linear and Nonlinear Estimation with Uncertain Observations," *Proc. American Control Conference*, Baltimore, MD, pp. 1429-1433, 1994.

[6] J. K. Tugnait, "Stability of Optimum Linear Estimators of Stochastic Signals in White Noise Multiplicative Noise," *IEEE Trans. Automatic Control*, vol. 26, pp. 757-761, 1981.

[7] E. Yaz, "Observer Design for Stochastic Parameter Systems," *Int. Journal Control*, vol. 46, pp. 1213-1217, 1987.

[8] E. Yaz, "Robustness of Stochastic Parameter Control and Estimation Schemes," *IEEE Trans. Automatic Control*, vol. 35, pp. 637-640, 1990.

[9] E. Yaz, "Full and Reduced Order Observer Design for Discrete Stochastic Bilinear Systems," *IEEE Trans. Automatic Control*, vol. 37, pp. 503-505, 1992.

[10] E. Yaz, "Observer Design for Discrete-Time Stochastic-Parameter Systems," in *Control and Dynamic Systems*, edited by C. T. Leondes, vol. 73, pp. 121-158, Academic Press, San Diego, CA, 1995.

[11] S. Boyd, L. El Ghaoui, E. Feiron and V. Balakrishnan, *LMIs in System and Control Theory*, SIAM: Philadelphia, PA, 1994.

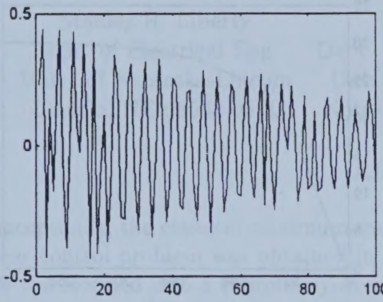


Figure 1 The measurement signal.

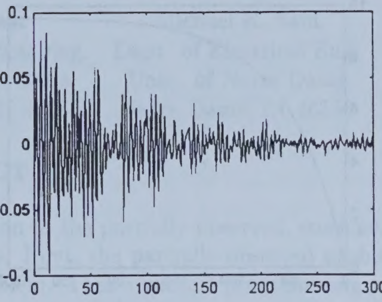


Figure 2 The error $\mathscr{E}\{y_k - \hat{y}_k\}$ for the first estimator.

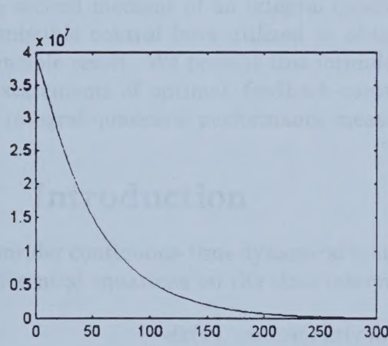


Figure 3 The error bound, $\mathscr{E}\{x_k^T P_1 x_k + e_k^T P_2 e_k\}$, for the first estimator.

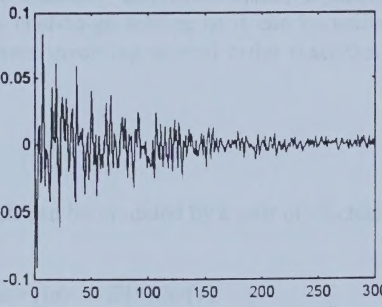
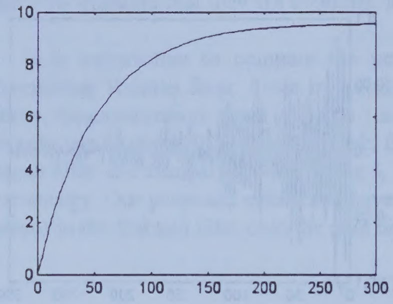


Figure 4 The error $\mathscr{E}\{y_k - \hat{y}_k\}$ for the second estimator.



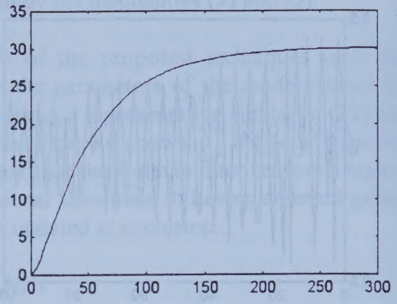


Figure 5 The error bound, $\sum_{i=0}^{k} \mathscr{E}\{\|e_k\|^2\}$, for the second estimator.

Figure 6 The error bound, $\sum_{i=0}^{k} \mathscr{E}\{\|e_k\|^2\}$, using the Kalman filter.

CONTROL ACCESS TO PERFORMANCE MEASURE STATISTICS: A SECOND MOMENT CASE *

Stanley R. Liberty
Dept. of Electrical Eng.
Univ. of Nebraska-Lincoln
Lincoln, NE 68588-0511

Peter Quast
Dept. of Electrical Eng.
Univ. of Notre Dame
Notre Dame, IN 46556

Michael K. Sain
Dept. of Electrical Eng.
Univ. of Notre Dame
Notre Dame, IN 46556

ABSTRACT

Conceptually, the classical minimum-mean solution to the partially-observed, stochastic-linear-control problem was obtained in two steps. First, the partially-observed problem was transformed into a completely-observed problem. Then dynamic programming was applied to the reformulated problem to obtain an optimal feedback control. The reformulated problem statement is sometimes called an accessible state formulation. Up to this time, it appears that no exact solution to partially-observed problems has surfaced for the case of control goals expressed in terms of finite numbers of higher order statistics of the performance measure. One reason for this is that no reformulation of higher-order performance measure statistics has been developed in terms of a truly accessible state — a variate which can be produced by a causal operation on the measured system output. This paper addresses a fundamental first question on the road to a complete understanding of such issues, by determining a causal accessible-state formulation for the second moment of an integral quadratic performance measure for the same class of admissible control laws utilized in obtaining the classical, minimum-mean, separation principle result. We present this formulation in a cost-to-go setting so it can be utilized in statements of optimal, feedback-control problems involving second order statistics of an integral-quadratic performance measure.

1 Introduction

Consider continuous-time dynamical systems which can be modeled by a pair of stochastic differential equations on the time interval $[t_0, t_f]$:

$$dx(\tau) = A(\tau)x(\tau)d\tau + b(\tau, u(\tau))d\tau + E(\tau)dw(\tau) \tag{1}$$

$$d\eta(\tau) = C(\tau)x(\tau)d\tau + dw_2(\tau) \tag{2}$$

where x is the state of the system, u is the control, and η is the "accumulated" measurement. The state initial condition $x(t_0)$ is a Gaussian distributed random variable such that $\mathcal{E}\{x(t_0)\} = \overline{x}_0$ and $\mathcal{E}\{[x(t_0) - \mathcal{E}\{x(t_0)\}][x(t_0) - \mathcal{E}\{x(t_0)\}]^T\} = \Sigma_0$. Let $w(\tau)$ and $w_2(\tau)$, $\tau \in [t_0, t_f]$ be stationary Wiener processes independent of each other and $x(t_0)$ with correlation of increments $\mathcal{E}\{(w(\tau) - w(\xi))(w(\tau) - w(\xi))^T\} = W|\tau - \xi|$ and $\mathcal{E}\{(w_2(\tau) - w_2(\xi))(w_2(\tau) - w_2(\xi))^T\} = W_2|\tau - \xi|$ respectively.

---

*The work of the first author was supported by the University of Nebraska Faculty Development Fellowship Leave Program. The work of the second and third authors was supported by the National Science Foundation under Grants CMS 95-00301 and 95-28083. The third author also wishes to acknowledge the support of the Frank M. Freimann Chair in Electrical Engineering at the University of Notre Dame.

For the purpose of measuring the performance of this system under selected control actions, we select the performance measure:

$$J = \int_{t_0}^{t_f} \Big[x^T(\tau)Q(\tau)x(\tau) + u^T(\tau)R(\tau)u(\tau)\Big]d\tau \tag{3}$$

We have not included a "terminal penalty" term in the performance measure for the sake of brevity and clarity in this presentation. However, inclusion of such a term poses no additional technical difficulties. Also, technical details on these equations and the class of admissible controls are not essential to our introductory remarks and are presented in the next section.

It is desirable, for obvious practical reasons, to restrict selection of controls to those which are non-anticipative functions of the measured system output; i.e. causal output-feedback control laws. Application of optimization theory to the issue of selecting a causal feedback control law must involve statistical criteria related to the random variable J, which we will frequently call the "cost". The most studied and most well known problem of this type is the minimum-mean Linear Quadratic Gaussian (LQG) problem where the cost mean is minimized by linear memoryless feedback of the Kalman-Bucy filtered estimate of the system state, a result which first appeared in the "separation theorem" paper by Wonham [15].

However, there are many other possible LQG problem statements involving higher order statistics of J. For example, in the 1960s, Sain conducted the first research on stochastic optimal control problems involving second order statistics of J, specifically the variance of J [5–7]. Since then Sain and colleagues have researched many aspects of minimal cost variance (MCV) control [1, 3, 8–12]. Sain's earliest efforts included the open-loop MCV solution and more recently the doctoral work of Won [12] included the solution of the continuous-time complete-state-feedback MCV problem. However, at the time of this writing, the complete solution to an output-feedback MCV control problem has not yet been reported. To appreciate why the output-feedback MCV problem has been so challenging consider the optimization approach.

A theory of optimization which is most compatible with enforcement of the feedback control law requirement is the Hamilton-Jacobi-Bellman (HJB) theory [2]. The HJB approach to the MCV control problem first requires that we "imbed" the original problem statement in a "cost-to-go" formulation involving the conditional mean and variance (or second moment) of J, expressed explicitly as functions of the present time and conditioned on the accumulated measurements up to the present time. Second, since in the above system only incomplete information of the system state is available, (the "partially observed problem" [2]), at any time through the output measurements, we must reformulate the conditional statistics of J in terms of an "accessible state" which can be physically produced by a causal operation on the measured system output. Finally, an optimization problem must be stated and the associated HJB equation derived and solved. Indeed, these were the steps taken in solving the minimum mean problem in [15]. However, for MCV problems involving an integral cost function such as that given by Equation 3, all of these steps are made more challenging by the fact that the second order statistics of J are two-dimensional integral expressions. More than one avenue of approach can be followed in transforming such integral expressions into statistics of one type or another. For example, in [4]), the integrals transform quite naturally to expressions which involve the fixed-interval smoothed estimate of the system state. Such statistics, however, are not physically realizable for finite-dimensional, causal feedback control. Instead, it is necessary to form an causal approximation to the overall expression for design purposes by omitting certain portions of it. A similar approach to approximating the precise statistical expression by neglecting a term or two was investigated in [12]. While examples of such simplified controllers have turned out to be interesting, the basic question of accessible state formulation in terms of causal, finite-dimensional measurement processing has remained. Moreover, in order to make an appropriate assessment of the utility of such simplifications, it is necessary to have in hand an exact solution.

This paper, therefore, develops an accessible-state formulation of the second conditional moment of J which is expressed in terms of physically realizable variates: namely the filtered and fixed-point smoothed estimates of the system state. In achieving this representation, the results herein overcome one of the key roadblocks seen in [4, 12], and set the stage for future development of output-feedback, cost-statistic controllers, a new and rich family of controllers in the broad spirit of risk sensitive control and related viewpoints [13].

2 Preliminaries

We will begin by presenting definitions that will be necessary for our development throughout the paper.

2.1 Definitions and Notation

Take T to be the compact interval $[t_0, t_f] \subset \mathbb{R}$. Let $\mathcal{Q}_n = (t_0, t_f) \times \mathbb{R}^n$. Let $\overline{\mathcal{Q}}_n$ be the closure of $\mathcal{Q}_n$. Let $x(\tau) \in \mathbb{R}^n$ represent the state of the system and $u(\tau)$ the control for the system at time τ taken from a convex, compact set $U \subset \mathbb{R}^m$. Let $\eta(\tau) \in \mathbb{R}^o$ represent the accessible accumulated measured output of the system. Consider a function f over a domain D defined $f : D \to \mathbb{R}^q$. Then let $C(D)$ denote the class of functions which are continuous on D (or an open set containing D if D is not open) and $C^i(D)$ the class of functions for which the i$^{\text{th}}$ derivative and all lower derivatives are continuous on D. We say that a real valued function $f(t,x)$ satisfies a *polynomial growth condition* on $\mathcal{Q}_n$ if for some constants c and k, $|f(t,x)| \leq c(1+|x|^k)$ when $(t,x) \in \mathcal{Q}_n$. Let $C_p^i(\mathcal{Q}_n)$ denote the class of functions $f(t,x) \in C^i(\mathcal{Q}_n)$ which satisfy a polynomial growth condition on $\mathcal{Q}_n$. The notation $g_{[a,b]} \triangleq g\big|[a,b]$ shall be used to refer to the restriction of the function $g(\tau)$, $\tau \in [t_0, t_f]$ to the subinterval $[a,b] \subseteq [t_0, t_f]$. The transpose of a matrix M will be denoted M^T.

2.2 Assumptions Regarding Equations 1, 2 and 3

The system and performance measure described in Equations 1 - 3 are assumed to have the following properties:

Assumption 2.1 *Assume* $A(\cdot)$, $C(\cdot)$, $E(\cdot)$, $Q(\cdot)$ *and* $R(\cdot)$ *are elements of* $C^1([t_0, t_f])$, $Q(\cdot)$ *positive semidefinite,* $R(\cdot)$ *positive definite and* $b(\cdot, u(\cdot)) \in C^1(\overline{\mathcal{Q}}_n \times U)$. *For some constants* c_0, c_1 *and* c_2 *and* $\tau \in [t_0, t_f]$, *assume* $b(\tau, u(\tau)) \leq c_0\,(1 + ||x(\tau)||)$, $\left(||A(\tau)|| + \left|\left|\frac{\partial b(\tau,u(\tau))}{\partial u}\right|\right|\right) \leq c_1$ *and* $||E(\tau)|| \leq c_2$.

Definition 2.1 *(Fleming) Define a set of mappings* $\Psi = \left\{\psi : [t_0, t_f] \times C([t_0, t_f]) \to U\right\}$ *which satisfies: 1) (Causality Condition) If* $f(\tau) = g(\tau)$ *for* $t_0 \leq \tau \leq t$, *then* $\psi(t, f) = \psi(t, g)$ *2) (Uniform Lipschitz Condition) For each* $\psi \in \Psi$, *there exists a constant* c_1 *such that* $||\psi(t,f) - \psi(t,g)|| < c_1\,||f - g||$, $\forall t \in [t_0, t_f]$, $\forall f, g \in C([t_0, t_f])$ *(where* $||\cdot||$ *denotes the* sup norm *in* $C([t_0, t_f])$*) 3)* $\psi(\cdot, \cdot)$ *is Borel measurable with respect to* $\overline{\mathcal{Q}}_n$ *4)*

$\psi(t,0)$ is bounded $\forall t \in [t_0, t_f]$. Define the class of **admissible controls** *$\mathcal{U}$ as functions of past accumulated measurements (nonanticipative) of the form $u(t) = \psi(t, \eta)$, $\psi \in \Psi$, $t \in [t_0, t_f]$. Observe that an admissible control $u(t)$ is $\mathcal{F}_t$ measurable.*

For each $u \in \mathcal{U}$, the conditions of Assumption 2.1 are sufficient for existence and uniqueness of a solution to stochastic differential Equations 1 and 2 in the Ito sense [2] and for assuring the validity of the mathematical devices used in developing the results presented in this paper.

2.3 Controlled/Uncontrolled Decomposition

Consider the state $x(\tau)$ as being composed of two components $x_1(\tau)$ and $x_2(\tau)$ such that:

$$dx_1(\tau) = A(\tau)x_1(\tau)dt + b(\tau, u(\tau))d\tau \qquad x_1(t_0) = 0 \tag{4}$$

$$dx_2(\tau) = A(\tau)x_2(\tau)dt + E(\tau)dw(\tau) \qquad x_2(t_0) = x(t_0) \tag{5}$$

Thus $x(\tau) = x_1(\tau) + x_2(\tau)$. Likewise, $d\eta(\tau) = C(\tau)\big(x_1(\tau) + x_2(\tau)\big)d\tau + dw_2(\tau)$. Define a process $\eta_2(\tau)$ by:

$$d\eta_2(\tau) \triangleq C(\tau)x_2(\tau)d\tau + dw_2(\tau) \tag{6}$$

2.4 Additional Notation and Terminology

Definition 2.2 *$\mathcal{F}_\xi$ is the minimal σ algebra generated by the accumulated measured output process $\eta(\tau) : \tau \in [t_0, \xi]$. Observe that $\mathcal{F}_{t_0}$ is empty since at $\tau = t_0$ there are no accumulated measurements. The notation $\mathcal{E}_\gamma\{\,.\,\}$ represents the conditional expectation taken with respect to the sigma algebra $\mathcal{F}_\gamma$,* i.e. *$\mathcal{E}_\gamma\{\,.\,\} \triangleq \mathcal{E}\left\{\,.\,\middle|\,\mathcal{F}_\gamma\right\}$.*

We will also use the following notation for objects related to the least squares filtered and smoothed estimates:

Definition 2.3 *For $\xi < \gamma$, define $\hat{x}(\xi|\xi) \triangleq \mathcal{E}\{x(\xi)|\,\mathcal{F}_\xi\}$, $\hat{x}(\xi|\gamma) \triangleq \mathcal{E}\{x(\xi)|\,\mathcal{F}_\gamma\}$, $\tilde{x}(\xi|\xi) \triangleq x(\xi) - \hat{x}(\xi|\xi)$, $\tilde{x}(\xi|\gamma) \triangleq x(\xi) - \hat{x}(\xi|\gamma)$, $K(\xi,\gamma) \triangleq \mathcal{E}\Big\{[x(\xi) - \hat{x}(\xi|\xi)]\,[x(\gamma) - \hat{x}(\gamma|\gamma)]^T\Big\}$, $\Gamma(\xi,\gamma;\theta) \triangleq \mathcal{E}\Big\{[x(\xi) - \hat{x}(\xi|\theta)]\,[x(\gamma) - \hat{x}(\gamma|\theta)]^T\Big\}$, $\Gamma(\xi,\gamma|\theta) \triangleq \mathcal{E}\Big\{[x(\xi) - \hat{x}(\xi|\theta)]\,[x(\gamma) - \hat{x}(\gamma|\theta)]^T\Big|\,\mathcal{F}_\theta\Big\}$, $\Sigma(\xi) \triangleq K(\xi,\xi)$. Also define the innovation process $\hat{w}(\tau)$ as $d\hat{w}(\tau) \triangleq C(\tau)\tilde{x}(\tau|\tau)d\tau + dw_2(\tau)$. The innovation process is a Wiener process with correlation of increments $\mathcal{E}\Big\{(\hat{w}(\tau) - \hat{w}(\xi))(\hat{w}(\tau) - \hat{w}(\xi))^T\Big\} = W_2|\tau - \xi|$. In the sequel, we will designate $\mathcal{G}(t) \triangleq C^T(t)W_2^{-1}C(t)$.*

Definition 2.4 *$\mathcal{F}^2_\xi$ is the minimal σ algebra generated by the process $\eta_2(\tau) : \tau \in [t_0, \xi]$. Also define $\hat{x}_2(\xi|\xi) \triangleq \mathcal{E}\Big\{x_2(\xi)|\,\mathcal{F}^2_\xi\Big\}$, $\hat{x}_2(\xi|\gamma) \triangleq \mathcal{E}\Big\{x_2(\xi)|\,\mathcal{F}^2_\gamma\Big\}$, $\tilde{x}_2(\xi|\xi) \triangleq x_2(\xi) - \hat{x}_2(\xi|\xi)$, $\tilde{x}_2(\xi|\gamma) \triangleq x_2(\xi) - \hat{x}_2(\xi|\gamma)$.*

3 The Conditional Second Moment of J

The second (unconditional) moment of J is given by

$$\mathcal{E}\left\{J^2(u_{[t_0,t_f]})\right\} =$$
$$\mathcal{E}\left\{\int_{t_0}^{t_f}\int_{t_0}^{t_f}\left[x^T(\gamma)Q(\gamma)x(\gamma)+u^T(\gamma)R(\gamma)u(\gamma)\right]\left[x^T(\xi)Q(\xi)x(\xi)+u^T(\xi)R(\xi)u(\xi)\right]d\xi d\gamma\right\}$$

However, application of HJB theory to the solution of optimization problems involving the second moment will require that these problems be imbedded in a "cost-to-go" formulation. There are two requirements for such a formulation. First, in the second moment expression, t_0 must be replaced by any t in the interval $[t_0, t_f]$. Second, the expectation operator $\mathcal{E}$ must be replaced by the conditional expectation, $\mathcal{E}_t$. Since the sigma algebra $\mathcal{F}_{t_0}$ is empty, it is readily seen that at $t = t_0$, $\mathcal{E}_{t_0}\left\{J^2(u_{[t_0,t_f]})\right\} = \mathcal{E}\left\{J^2(u_{[t_0,t_f]})\right\}$. Also, since the integrand in the double integral expression of the second moment is scalar and symmetric in its time arguments, it is easy to show that the double integral can be replaced by what is commonly called an "iterated integral". Therefore we have

$$\mathcal{E}_t\left\{J^2(u_{[t,t_f]})\right\} =$$
$$2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[x^T(\gamma)Q(\gamma)x(\gamma)+u^T(\gamma)R(\gamma)u(\gamma)\right]\left[x^T(\xi)Q(\xi)x(\xi)+u^T(\xi)R(\xi)u(\xi)\right]d\xi d\gamma\right\} \tag{7}$$

The main result of this paper can now be stated as

Theorem 3.1 *For admissible controls $u_{[t_0,t_f]} \in \mathcal{U}$, the cost-to-go formulation for the second moment of J can be reexpressed in terms of the filtered and fixed-point smoothed estimates of the system state as:*

$$\begin{aligned}
&\mathcal{E}_t[J^2(u_{[t,t_f]})] = \\
&\quad 2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\hat{x}^T(\gamma|\gamma)Q(\gamma)\hat{x}(\gamma|\gamma)+u^T(\gamma)R(\gamma)u(\gamma)\right]\right. \\
&\qquad\qquad\qquad \left.\cdot\left[\hat{x}^T(\xi|\gamma)Q(\xi)\hat{x}(\xi|\gamma)+u^T(\xi)R(\xi)u(\xi)\right]d\xi d\gamma\right\} \\
&\quad +8\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\hat{x}^T(\gamma|\gamma)Q(\gamma)K(\gamma,\xi)Q(\xi)\hat{x}(\xi|\gamma)d\xi d\gamma\right\} \\
&\quad +Tr\left[\int_t^{t_f}Q(\gamma)\Sigma(\gamma)d\gamma\right]Tr\left[\int_t^{t_f}Q(\xi)\Sigma(\xi)d\xi\right] \\
&\quad +4Tr\left[\int_t^{t_f}\int_t^{\gamma}Q(\gamma)K(\gamma,\xi)Q(\xi)K(\xi,\gamma)d\xi d\gamma\right] \\
&\quad +2\mathcal{E}_t\left\{\int_t^{t_f}\left[\hat{x}^T(\gamma|\gamma)Q(\gamma)\hat{x}(\gamma|\gamma)+u^T(\gamma)R(\gamma)u(\gamma)\right]d\gamma\right\}Tr\left[\int_t^{t_f}Q(\xi)\Sigma(\xi)d\xi\right] \\
&\quad -2\mathcal{E}_t\left\{\int_t^{t_f}\left[\hat{x}^T(\gamma|\gamma)Q(\gamma)\hat{x}(\gamma|\gamma)+u^T(\gamma)R(\gamma)u(\gamma)\right]\right. \\
&\qquad\qquad\qquad \left.\cdot\int_t^{\gamma}Tr\left[Q(\xi)\int_\xi^{\gamma}K(\xi,\rho)\mathcal{G}(\rho)K(\rho,\xi)d\rho\right]d\xi d\gamma\right\}
\end{aligned}$$

4 Lemmas

The following lemmas are utilized in the proof of the above theorem. We have omitted the proofs for brevity.

Lemma 4.1 *For the system described by Equations 1 and 2 with $u \in \mathcal{U}$, the smoothed estimate and filtered estimate of the system state $x(\xi)$, $\xi \leq \gamma$ are related by:*

$$\widehat{x}(\xi|\gamma) \;=\; \widehat{x}(\xi|\xi) + \int_\xi^\gamma K(\xi,\tau)C^T(\tau)W_2^{-1}d\widehat{w}(\tau) \tag{8}$$

Define the integral $L(\xi,\gamma;\widehat{w}) \triangleq \int_\xi^\gamma K(\xi,\tau)C^T(\tau)W_2^{-1}(\tau)d\widehat{w}(\tau)$.

Lemma 4.2 *For system described by Equation 1 and integral $L(\xi,\gamma;\widehat{w})$ defined above, the following expressions hold:*

1. *For $t \leq \xi$: $\mathcal{E}_t\big\{L(\xi,\gamma;\widehat{w})\big\} = 0$*
2. *For overlapping intervals $[\xi,\gamma]$ and $[\alpha,\theta]$, $t \leq max(\alpha,\xi)$:*
$$\mathcal{E}_t\big\{L(\xi,\gamma;\widehat{w})L^T(\alpha,\theta;\widehat{w})\big\} \;=\; \int_{\max(\alpha,\xi)}^{\min(\gamma,\theta)} K(\xi,\tau)\mathcal{G}(\tau)K(\tau,\alpha)d\tau$$
3. *For $\alpha \leq \xi$, $t \leq \xi$:*
$$\mathcal{E}_t\left\{L(\xi,\gamma;\widehat{w})\widehat{x}^T(\alpha|\alpha)\right\} = 0$$

Lemma 4.3 *For the system given in Equation 1, the smoothed error covariance $\Gamma(\xi,\theta;\gamma)$, $max(\xi,\theta) \leq \gamma$ can be expressed as:*

$$\Gamma(\xi,\theta;\gamma) \;=\; K(\xi,\theta) - \int_{\max(\xi,\theta)}^{\gamma} K(\xi,\tau)\mathcal{G}(\tau)K(\tau,\theta)d\tau$$

5 Proof of Theorem 3.1

Proof: Begin with the cost-to-go formulation for the second moment of J given by Equation 7 and substitute for $x(\gamma)$ and $x(\xi)$ in terms of filtered or smoothed estimates conditioned on $\mathcal{F}_\gamma$ and their associated error ($\xi \leq \gamma$, $x(\gamma) = \widehat{x}(\gamma|\gamma) + \widetilde{x}(\gamma|\gamma)$, $x(\xi) = \widehat{x}(\xi|\gamma) + \widetilde{x}(\xi|\gamma)$). We get:

$$\begin{aligned}
&\mathcal{E}_t[J^2(u_{[t,t_f]})] \\
&= 2\mathcal{E}_t\Big\{\int_t^{t_f}\int_t^\gamma \big[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma) + u^T(\gamma)R(\gamma)u(\gamma)\big] \\
&\qquad\qquad \cdot\big[\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma) + u^T(\xi)R(\xi)u(\xi)\big]\,d\xi d\gamma\Big\} && (a) \\
&\quad +2\mathcal{E}_t\Big\{\int_t^{t_f}\int_t^\gamma \big[\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\big]\,d\xi d\gamma\Big\} && (b) \\
&\quad +4\mathcal{E}_t\Big\{\int_t^{t_f}\int_t^\gamma \big[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma)\widehat{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\big]\,d\xi d\gamma\Big\} && (g_1) \\
&\quad +2\mathcal{E}_t\Big\{\int_t^{t_f}\int_t^\gamma \big[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\big]\,d\xi d\gamma\Big\} && (d_2) \\
&\quad +8\mathcal{E}_t\Big\{\int_t^{t_f}\int_t^\gamma \big[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma)\big]\,d\xi d\gamma\Big\} && (c) \\
&\quad +4\mathcal{E}_t\Big\{\int_t^{t_f}\int_t^\gamma \big[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma)\big]\,d\xi d\gamma\Big\} && (g_2) \\
&\quad +4\mathcal{E}_t\Big\{\int_t^{t_f}\int_t^\gamma \big[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\big]\,d\xi d\gamma\Big\} && (f_1)
\end{aligned}$$

$$+2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma)\right]d\xi d\gamma\right\} \qquad (d_1)$$

$$+4\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widehat{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\right]d\xi d\gamma\right\} \qquad (f_2)$$

$$+2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)u^T(\gamma)R(\gamma)u(\gamma)\right]d\xi d\gamma\right\} \qquad (e_1)$$

$$+4\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)u^T(\gamma)R(\gamma)u(\gamma)\right]d\xi d\gamma\right\} \qquad (h_1)$$

$$+4\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[u^T(\gamma)R(\gamma)u(\gamma)\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\right]d\xi d\gamma\right\} \qquad (h_2)$$

$$+2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[u^T(\gamma)R(\gamma)u(\gamma)\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\right]d\xi d\gamma\right\} \qquad (e_2)$$

Note: The letters in parenthesis after each term will be a convention used throughout the sequel to track terms. Since $u(\gamma)$ and $u(\xi)$ are $\mathcal{F}_\gamma$ measurable by assumption of $u(\cdot) \in \mathcal{U}$ and $\widehat{x}(\gamma|\gamma)$, $\widehat{x}(\xi|\gamma)$, $\xi \leq \gamma$ are $\mathcal{F}_\gamma$ measurable by definition [14], terms g_1, g_2, h_1 and h_2 can be shown to be equal to 0. For example, examining term g_1:

$$\begin{aligned} g_1 &= 4\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\right]d\xi d\gamma\right\} \\ &= 4\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma)\widehat{x}^T(\xi|\gamma)Q(\xi)\mathcal{E}_\gamma\left\{\widetilde{x}(\xi|\gamma)\right\}\right]d\xi d\gamma\right\} \end{aligned}$$

However, $\mathcal{E}_\gamma\left\{\widetilde{x}(\xi|\gamma)\right\} = 0$ and thus $g_1 = 0$. Terms f_1 and f_2 can also be shown to be equal to 0. Examining term f_1:

$$\begin{aligned} f_1 &= 4\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\right]d\xi d\gamma\right\} \\ &= 4\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\mathcal{E}_\gamma\left\{\widetilde{x}_2(\gamma|\gamma)\widetilde{x}_2^T(\xi|\gamma)Q(\xi)\widetilde{x}_2(\xi|\gamma)\right\}\right]d\xi d\gamma\right\} \end{aligned}$$

since $\widehat{x}(\gamma|\gamma)$ is $\mathcal{F}_\gamma$ measurable. Since $\mathcal{E}_\gamma\left\{\widetilde{x}_2(\xi|\gamma)\right\} = 0$ and $\mathcal{E}_\gamma\left\{\widetilde{x}_2(\gamma|\gamma)\right\} = 0$, it can be shown that $\mathcal{E}_t\left\{\widetilde{x}_2(\gamma|\gamma)\widetilde{x}_2^T(\xi|\gamma)Q(\xi)\widetilde{x}_2(\xi|\gamma)\right\} = 0$. Thus $f_1 = 0$. The same argument follows for term f_2. Considering terms d_1, d_2, e_1 and e_2, first examine term d_1:

$$\begin{aligned} d_1 &= 2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma)\right]d\xi d\gamma\right\} \\ &= 2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\mathcal{E}_\gamma\left\{\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\right\}\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma)\right]d\xi d\gamma\right\} \\ &= 2\mathcal{E}_t\left\{\int_t^{t_f} Tr\left[Q(\gamma)\Sigma(\gamma)\right]\int_t^{\gamma}\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma)d\xi d\gamma\right\} \end{aligned}$$

since $\widehat{x}(\xi|\gamma)$ is $\mathcal{F}_\gamma$ measurable by definition. Using the same argument for term e_1, we get:

$$\begin{aligned} d_1 + e_1 &\triangleq de_1 \\ &= 2\mathcal{E}_t\left\{\int_t^{t_f} Tr\left[Q(\gamma)\Sigma(\gamma)\right]\left[\int_t^{\gamma}\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma) + u^T(\xi)R(\xi)u(\xi)\right]d\xi d\gamma\right\} \end{aligned}$$

Similarly, for term d_2:

$$\begin{aligned} d_2 &= 2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widetilde{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma)\mathcal{E}_\gamma\left\{\widetilde{x}^T(\xi|\gamma)Q(\xi)\widetilde{x}(\xi|\gamma)\right\}\right]d\xi d\gamma\right\} \\ &= 2\mathcal{E}_t\left\{\int_t^{t_f}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma)\right]\int_t^{\gamma} Tr\left[Q(\xi)\Gamma(\xi,\xi;\gamma)\right]d\xi d\gamma\right\} \end{aligned}$$

Using the same argument for term e_2, we get:

$$\begin{aligned} d_2 + e_2 &\triangleq de_2 \\ &= 2\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma) + u^T(\gamma)R(\gamma)u(\gamma)\right] Tr\left[\int_t^{\gamma} Q(\xi)\Gamma(\xi,\xi;\gamma)\right] d\xi d\gamma\right\} \end{aligned}$$

Examining term b, we note that it is independent of control. It consists of 4 zero mean, jointly Gaussian variates. Using Lemma 4.3, it can be shown that:

$$\begin{aligned} b &= 2\int_t^{t_f}\int_t^{\gamma} Tr\left[Q(\gamma)\Sigma(\gamma)\right] Tr\left[Q(\xi)\Gamma(\xi,\xi;\gamma)\right] d\xi d\gamma \\ &\quad +4Tr\left[\int_t^{t_f}\int_t^{\gamma} Q(\gamma)K(\gamma,\xi)Q(\xi)K(\xi,\gamma)d\xi d\gamma\right] \end{aligned}$$

Examining term c:

$$\begin{aligned} c &= 8\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widetilde{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma)d\xi d\gamma\right\} \\ &= 8\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\widehat{x}^T(\gamma|\gamma)Q(\gamma)\mathcal{E}_{\gamma}\left\{\widetilde{x}(\gamma|\gamma)\widetilde{x}^T(\xi|\gamma)\right\} Q(\xi)\widehat{x}(\xi|\gamma)d\xi d\gamma\right\} \\ &= 8\mathcal{E}_t\left\{\int_t^{t_f}\int_t^{\gamma}\widehat{x}^T(\gamma|\gamma)Q(\gamma)K(\gamma,\xi)Q(\xi)\widehat{x}(\xi|\gamma)d\xi d\gamma\right\} \end{aligned}$$

Substituting Equation 8 in the expression for de_1:

$$\begin{aligned} de_1 &= 2\mathcal{E}_t\left\{\int_t^{t_f} Tr\left[Q(\gamma)\Sigma(\gamma)\right]\left[\int_t^{\gamma}\widehat{x}^T(\xi|\gamma)Q(\xi)\widehat{x}(\xi|\gamma) + u^T(\xi)R(\xi)u(\xi)\right] d\xi d\gamma\right\} \\ &= 2\mathcal{E}_t\left\{\int_t^{t_f} Tr\left[Q(\gamma)\Sigma(\gamma)\right]\int_t^{\gamma}\left[\widehat{x}^T(\xi|\xi)Q(\xi)\widehat{x}(\xi|\xi) + u^T(\xi)R(\xi)u(\xi)\right] d\xi d\gamma\right\} \quad (de_{11}) \\ &\quad +4\int_t^{t_f} Tr\left[Q(\gamma)\Sigma(\gamma)\right] Tr\left[\int_t^{\gamma} Q(\xi)\mathcal{E}_t\left\{L(\xi,\gamma;d\widehat{w})\widehat{x}^T(\xi|\xi)\right\} d\xi\right] d\gamma \quad (de_{13}) \\ &\quad +2\int_t^{t_f} Tr\left[Q(\gamma)\Sigma(\gamma)\right] Tr\left[\int_t^{\gamma} Q(\xi)\mathcal{E}_t\left\{L(\xi,\gamma;d\widehat{w})L^T(\xi,\gamma;d\widehat{w})\right\} d\xi\right] d\gamma \quad (de_{12}) \end{aligned}$$

Since $\mathcal{E}_t\left\{L(\xi,\gamma;d\widehat{w})\widehat{x}^T(\xi|\xi)\right\} = 0$ by Lemma 4.2, term $de_{13} = 0$. Also, by Lemma 4.2, $\mathcal{E}_t\left\{L(\xi,\gamma;d\widehat{w})L^T(\xi,\gamma;d\widehat{w})\right\} = \int_{\xi}^{\gamma} K(\xi,\tau)\mathcal{G}(\tau)K(\tau,\xi)d\tau$. Combining terms b and de_{12} and applying Lemma 4.3:

$$\begin{aligned} b + de_{12} &= 2\int_t^{t_f}\int_t^{\gamma} Tr\left[Q(\gamma)\Sigma(\gamma)\right] Tr\left[Q(\xi)\left[\Gamma(\xi,\xi;\gamma) + \int_{\xi}^{\gamma} K(\xi,\tau)\mathcal{G}(\tau)K(\tau,\xi)d\tau\right]\right] d\xi d\gamma \\ &\quad +4Tr\left[\int_t^{t_f}\int_t^{\gamma} Q(\gamma)K(\gamma,\xi)Q(\xi)K(\xi,\gamma)d\xi d\gamma\right] \\ &= Tr\left[\int_t^{t_f} Q(\gamma)\Sigma(\gamma)d\gamma\right] Tr\left[\int_t^{t_f} Q(\xi)\Sigma(\xi)d\xi\right] \\ &\quad +4Tr\left[\int_t^{t_f}\int_t^{\gamma} Q(\gamma)K(\gamma,\xi)Q(\xi)K(\xi,\gamma)d\xi d\gamma\right] \end{aligned}$$

Considering term de_2 in view of Lemma 4.3:

$$\begin{aligned} de_2 &= 2\mathcal{E}_t\left\{\int_t^{t_f}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma) + u^T(\gamma)R(\gamma)u(\gamma)\right]\int_t^{\gamma} Tr\left[Q(\xi)\Sigma(\xi)\right] d\xi d\gamma \quad (de_{21}) \right. \\ &\quad -2\mathcal{E}_t\left\{\int_t^{t_f}\left[\widehat{x}^T(\gamma|\gamma)Q(\gamma)\widehat{x}(\gamma|\gamma) + u^T(\gamma)R(\gamma)u(\gamma)\right]\right. \\ &\quad\quad \left. \cdot\int_t^{\gamma} Tr\left[Q(\xi)\int_{\xi}^{\gamma} K(\xi,\rho)\mathcal{G}(\rho)K(\rho,\xi)d\rho\right] d\xi d\gamma\right\} \quad (de_{22}) \end{aligned}$$

Adding de_{11} and de_{21}:

$$\begin{aligned} de_{11} + de_{21} &= 2\mathcal{E}_t \left\{ \int_t^{t_f} Tr\left[Q(\gamma)\Sigma(\gamma)\right] \int_t^{\gamma} \left[\hat{x}^T(\xi|\xi)Q(\xi)\hat{x}(\xi|\xi) + u^T(\xi)R(\xi)u(\xi)\right] d\xi d\gamma \right\} \\ &\quad + 2\mathcal{E}_t \left\{ \int_t^{t_f} \left[\hat{x}^T(\gamma|\gamma)Q(\gamma)\hat{x}(\gamma|\gamma) + u^T(\gamma)R(\gamma)u(\gamma)\right] \int_t^{\gamma} Tr\left[Q(\xi)\Sigma(\xi)\right] d\xi d\gamma \right\} \\ &= 2\mathcal{E}_t \left\{ \int_t^{t_f} \left[\hat{x}^T(\gamma|\gamma)Q(\gamma)\hat{x}(\gamma|\gamma) + u^T(\gamma)R(\gamma)u(\gamma)\right] d\gamma \right\} Tr\left[\int_t^{t_f} Q(\xi)\Sigma(\xi) d\xi \right] \end{aligned}$$

Gathering terms a, b_1, b_2, c, $(de_{21} + de_{11})$ and de_{22}, we arrive at the desired result.

Q.E.D.

6 Conclusions

We believe the development of the second conditional moment expression presented in Theorem 3.1 represents a key step in the process of formulating MCV, or second moment, optimal feedback control problem statements. Although the two-integral complexity may appear to be formidable, we know this not to be the case based on Won's success with the complete- state-feedback MCV problem [10, 12]. On the other hand, the presence of the second variate, the fixed-point smoothed estimate of the state, raises new and interesting questions.

It is well known that the filtered estimate of the system state is a sufficient statistic for the conditional probability distribution of the system state, where the conditioning is on past accumulated observations [15]. Also the conditional mean of J can be expressed solely in terms of the filtered estimate. So, in the second moment case the fixed-point smoothed estimate appears to play a functional role arising from the two- integral nature of the second moment. It will be interesting to see if MCV or optimal second moment control laws couple to this variate. If so, these control laws will likely contain dynamics beyond those in the filtered-state estimator. It will also be interesting to see the implications, with regard to this variate, of further restrictions on the class of admissible control laws.

References

[1] Luis Cosenza. *On the Minimum Variance Control of Discrete-Time Systems.* PhD thesis, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, January 1969.

[2] Wendell H. Fleming and Raymond W. Rishel. *Deterministic and Stochastic Optimal Control.* Springer-Verlag, 1975.

[3] Stanley R. Liberty. *Characteristic Functions of LQG Control.* PhD thesis, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, August 1971.

[4] Stanley R. Liberty and Ronald C. Hartwig. Design-performance-measure statistics for stochastic linear control systems. *IEEE Transactions on Automatic Control*, AC-23(6):1085–1090, December 1978.

[5] Michael K. Sain. Relative costs of mean and variance control for a class of linear, noisy systems. In *Proceedings of the 3^{rd} Annual Allerton Conference on Circuit and System Theory*, pages 121–129, 1965.

[6] Michael K. Sain. Control of linear systems according to the minimal variance criterion - a new approach to the disturbance problem. *IEEE Transactions on Automatic Control*, AC-11(1):118–122, January 1966.

[7] Michael K. Sain. A sufficient condition for the minimum variance control of markov processes. In *Proceedings of the 4th Annual Allerton Conference on Circuit and System Theory*, pages 593–599, 1966.

[8] Michael K. Sain and Stanley R. Liberty. Performance measure densities for a class of LQG control systems. *IEEE Transactions on Automatic Control*, AC-16(5):431–439, October 1971.

[9] Michael K. Sain, Chang-Hee Won, and Billie F. Spencer Jr. Cumulant minimization and robust control. In Tyrone E. Duncan and Bozenna Pasik-Duncan, editors, *Lecture Notes in Control and Information Sciences*, pages 411–425, New York, 1992. Springer-Verlag.

[10] Michael K. Sain, Chang-Hee Won, and Billie F. Spencer Jr. Cumulants in risk-sensitive control: The full-state-feedback cost variance case. In *Proceedings IEEE Conference on Decision and Control*, pages 1036–1041, 1995.

[11] Chang-Hee Won. Cost variance control and risk sensitivity. Master's thesis, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, July 1992.

[12] Chang-Hee Won. *Cost Cumulants in Risk-Sensitive and Minimal Cost Variance Control*. PhD thesis, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, November 1994.

[13] Chang-Hee Won, Michael K. Sain, and Billie F. Spencer Jr. Relationships in modern stochastic control: Risk-sensitive and minimal cost variance control. In *Proceedings of the 34th Annual Allerton Conference on Communication, Control, and Computing*, pages 343–352, 1996.

[14] Eugene Wong and Bruce Hajek. *Stochastic Processes in Engineering Systems*. Springer-Verlag, 1985.

[15] W.M. Wonham. On the separation theorem of stochastic control. *SIAM Journal on Control*, 6(2):312–326, 1968.

Risk-Sensitive Queueing

Stefano P. Coraluppi and Steven I. Marcus
Electrical Engineering Department and
Institute for Systems Research
University of Maryland
College Park, MD 20742
stefano@isr.umd.edu, marcus@isr.umd.edu

Abstract

This paper has two objectives. The first is to introduce a unifying framework for risk-sensitive control of Markov Decision Processes (MDPs). The framework includes, as special cases, a number of approaches that have been considered in the literature. The second objective is to apply risk-sensitive criteria to the control of queueing systems. We generalize the uniformization methodology that relates continuous-time and discrete-time queueing models, and consider a simple discrete-time example.

Key Words: Stochastic Control, Risk-Sensitive Control, Markov Decision Processes, Queueing Systems

1 . Introduction

We consider the class of discrete-time MDPs with finite state space X, finite control space U, and finite observation space Y. The probability transition matrix $P(u)$ is defined by $P_{ij} = pr(x_{k+1} = j | x_k = i, u_k = u)$, and the observation matrix Q is defined by $Q_{ij} = pr(y_k = j | x_k = i)$. We define $c_k(x_k, u_k)$ to be the cost incurred by the system at time $k \geq 0$, given that it is in state $x_k \in X$ and that control $u_k \in U$ is used. If there is a finite horizon size N, there is a terminal cost $c_N(x_N)$.

The traditional stochastic control approach is to consider a *risk-neutral* objective given by

$$J(\mu, \pi_0) = E^{\mu,\pi_0}[\sum_k c_k(x, u)], \tag{1}$$

where μ is a non-anticipative policy and π_0 is the probability distribution on the states of the system at time $k = 0$. We use the terminology risk-neutral control to denote the fact that only the expected total cost incurred is considered. Higher order moments of the total cost are neglected. We can consider a generalization of (1) of the form

$$J^\gamma(\mu, \pi_0) = \frac{1}{\gamma} \log E^{\mu,\pi_0}[\exp(\gamma \sum_k c_k(x, u))]. \tag{2}$$

In the MDP context, this type of generalization was first considered by [HM72]. It is known as a *risk-sensitive* objective. For small γ, (2) takes the form

$$J^{\gamma}(\mu, \pi_0) \simeq E^{\mu,\pi_0}[\sum_k c_k(x, u)] + \frac{\gamma}{2} Var^{\mu,\pi_0}[\sum_k c_k(x, u)], \tag{3}$$

and in the limit $\gamma \to 0$, (2) reverts to the risk-neutral objective (1). The parameter γ allows one to incorporate an aversion or preference for risk, or variability in the cost incurred in the system's evolution. For $\gamma > 0$, we are penalized for variability in the cost incurred, so we say that we have a *risk-averse* objective. Key structural results for the synthesis of optimal risk-sensitive policies are due to [BJam] and [FGMar].

On the infinite horizon, in order for the objectives (1) and (2) to remain bounded, one can consider a discounted cost structure given by $c_k(\cdot, \cdot) = \beta^k c(\cdot, \cdot)$. It is well-known that with the risk-neutral objective (1), this leads to the optimality of a stationary policy. Unfortunately, with the risk-sensitive objective (2), in general there will not exist a stationary policy that is optimal (see [CS87]). The discount factor β leads to an *effective* risk-sensitive parameter of $\beta^k \gamma$ at time k.

The large risk-limit, $\gamma \to \infty$, of the risk-sensitive criterion (2) results in the following minimax criterion:

$$\bar{J}(\mu, \pi_0) = \sup_{\omega \in \Omega} \sum_k c_k(x_k, u_k), \tag{4}$$

where Ω is the set of trajectories of the form $(x_0, u_0, x_1, u_1, \ldots)$ that occur with non-zero probability under policy μ. Further details can be found in [Cor97]. Analogous results in the LQG setting can be found in [GD88]. An interesting observation is that, to the the extent that we are interested in worst-case system behavior, it is equivalent to consider a probabilistic or deterministic description of uncertainty.

2 A Generalized Decision-Making Framework

In the finite horizon, full state observations setting, we can define the value function corresponding to the risk-sensitive criterion (2) by

$$s^{\gamma}_{k,N}(i) := \min_{\mu} \frac{1}{\gamma} \log E^{\mu}[\exp(\gamma C_{k,N}) | x_k = i], \; i \in X, \tag{5}$$

where $C_{k,N} = \sum_{j=k}^{N-1} c_j(x_j, u_j) + c_N(x_N)$. The dynamic programming equations for (5) are given by

$$\begin{aligned} s^{\gamma}_{k,N}(i) &= \min_{u \in U}\{c_k(i, u) + \frac{1}{\gamma} \log[\sum_j P_{ij}(u) \exp(\gamma s^{\gamma}_{k+1,N}(j))]\}, \; k = 0, \ldots, N-1, & (6) \\ s^{\gamma}_{N,N}(i) &= c_N(i). & (7) \end{aligned}$$

In the small-risk limit, $\gamma \to 0$, (6), (7) revert to the usual risk-neutral dynamic programming equations. On the infinite horizon, we have

$$s^{\gamma}_{k}(i) = \min_{u \in U}\{c_k(i, u) + \frac{1}{\gamma} \log[\sum_j P_{ij}(u) \exp(\gamma s^{\gamma}_{k+1}(j))]\}, \; k = 0, \ldots. \tag{8}$$

If $c_k(\cdot,\cdot) = \beta^k c(\cdot,\cdot)$, it can be shown that time-shifted value functions are related as follows:

$$s_{k+1}^{\gamma}(\cdot) = \beta s_k^{\beta\gamma}(\cdot). \tag{9}$$

Equation (9) also reverts to a well-known relationship in the risk-neutral case:

$$s_{k+1}^{0}(\cdot) = \beta s_k^{0}(\cdot). \tag{10}$$

A more general set of optimality equations than (6), (7) can be defined as follows:

$$\begin{aligned} h_{k,N}^{\gamma}(i) &= \min_{u\in U}\{c_k(i,u) + \frac{\beta'}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma\beta'' h_{k+1,N}^{\gamma}(j))]\}, && (11)\\ h_{N,N}^{\gamma}(i) &= c_N(i). && (12)\end{aligned}$$

An interpretation for these optimality equations is that the value function at time k equals the cost incurred at time k, plus a (possibly discounted) contribution accounting for future costs. Note that if we set $\beta' = \beta'' = 1$, we revert to the classical risk-sensitive dynamic programing equations. If we set $\beta = \beta'' = 1$, we obtain the formulation that has been studies in a series of papers including [Por75] and [KP78], which we refer to as the *Porteus* formulation. A similar formulation in the LQG setting has been proposed recently in [HS95]. If we set $\beta = \beta' = 1$, we obtain the formulation introduced in [Eag75], which we refer to as the *Eagle* formulation.

On the infinite horizon, setting $c_k(\cdot,\cdot) = \beta^k c(\cdot,\cdot)$, the generalized optimality equation is given by

$$h_k^{\gamma}(i) = \min_{\mu}\{\beta^k c(i,u) + \frac{\beta'}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma\beta'' h_{k+1}^{\gamma}(j))]\},\ k = 0,\ldots. \tag{13}$$

Once again we obtain the classical, Porteus, and Eagle formulations as special cases of (13). A key feature of the generalized formulation (13) is that it is sufficient for one of β, β', and β'' to be less than 1, provided the others are set to 1, to insure boundedness of the value function h_k^{γ}. Thus, by setting either β' or β'' to be less than one, we can set $\beta = 1$. It can then be shown that $h_k^{\gamma}(\cdot) = h^{\gamma}(\cdot)$, that is we have a time-invariant value function, and furthermore there is a stationary policy that achieves the minimum in (13). This is proved by verifying that the *r.h.s.* of (13) defines a contraction mapping on $\Re_+^{|X|}$, where $\Re_+$ is the set of non-negative reals. It can further be shown that policy and value iteration techniques can be used to synthesize an optimal policy. See [Cor97] for details.

The nature of the discount factors β, β', and β'' can be better understood by considering the small-risk limit, $\gamma \to 0$, of (13). We obtain the following:

$$h_k^{0}(i) = \min_{u\in U}\{\beta^k c(i,u) + \beta'\beta''\sum_j P_{ij}(u) h_{k+1}^{0}(j)\},\ k = 0,\ldots. \tag{14}$$

Note that this optimality equation is more general than the risk-neutral dynamic programming equation. On the other hand, each of the three special cases of (13) that we have considered (classical, Porteus, Eagle) is equivalent to risk-neutral control in the small-risk limit.

An interesting consequence of introducing the additional discount parameters β' and β'' is that, unlike (6), (7), the equations (11), (12) are not dynamic programming equations. By this we mean that, in general, a policy $\mu^\star$ achieving the minimum on the *r.h.s.* of equations (11), (12) does not minimize a criterion of expected utility form. More precisely, in general there does not exist a $U : \Re_+ \to \Re_+$, such that the objective $E^\mu[U(\sum_k c_k(x_k, u_k))]$ is minimized by policy $\mu^\star$. The same comment applies to the infinite horizon optimality equation (13). This can be understood in light of the standard axiomatic formulation of Utility Theory, which can be found e.g. in [HS84].

3 Risk-Sensitive Uniformization

Models of interest for queueing applications may have either a continuous-time or a discrete-time formulation. We now introduce the standard, continuous-time formulation for queueing systems, and review the uniformization methodology by which we can transform the model to an equivalent discrete-time one. Further details can be found in [Ber95]. The transformation allows us to use the many tools available for discrete-time MDPs to address continuous-time queueing problems.

We will be interested in models for which we have a finite number of states. We assume that we have a known probability transition matrix $\tilde{P}(u)$, and that the transition time τ is exponentially distributed and depends on the current state i and control u, that is the probability density function of τ is

$$p(\tau) = \nu_i(u)\exp(-\nu_i(u)\tau),\ \tau \geq 0. \tag{15}$$

τ does not depend on earlier transition times, states, or controls. Finally, there exists a finite ν such that $\nu_i(u) \leq \nu$, for all states i and controls u.

The risk-neutral, infinite horizon, discounted cost criterion that we consider is given by

$$J(\mu, x_0) = E^{\mu, x_0}[\int_0^\infty \exp(-\alpha t)\tilde{c}(x(t), \mu(x(t)))dt],\ \alpha > 0, \tag{16}$$

where $x_0 = x(0)$. The state $x(t)$ and control $u(t) = \mu(x(t))$ are constant between transitions.

This model can be converted to an equivalent, discrete-time MDP with discount factor $\beta = \frac{\nu}{\alpha+\nu}$, transition probability matrix $P(u)$ given by

$$P_{ij}(u) = \begin{cases} \frac{\nu_i(u)}{\nu}\tilde{P}_{ij}, & \text{if } i \neq j, \\ 1 - \frac{\nu_i(u)}{\nu}, & \text{if } i = j, \end{cases} \tag{17}$$

and cost per stage given by

$$c_k(i, u) = \beta^k c(i, u), \tag{18}$$

where

$$c(i, u) = \frac{1}{\alpha + \nu}\tilde{c}(i, u). \tag{19}$$

Uniformization involves a re-scaling of the time axis, whereby the time of the k^{th} transition, t_k, is taken to be time k. The state of the system after the k^{th} transition is x_k, and the control is $u_k = \mu(x_k)$.

It is important to note that, besides the assumption about independent, exponentially distributed transition times, the form of the criterion (16) is key to the uniformization procedure. Indeed, we have

$$\begin{aligned} J(\mu, x_0) &= E^{\mu_0,x_0}[\int_0^\infty \exp(-\alpha t)\tilde{c}(x(t), \mu(x(t)))dt] \\ &= \sum_{k=0}^{\infty} E^{\mu_0,x_0}[\int_{t_k}^{t_{k+1}} \exp(-\alpha t)\tilde{c}(x(t), \mu(x(t)))dt] \\ &= \sum_{k=0}^{\infty} E[\int_{t_k}^{t_{k+1}} \exp(-\alpha t)dt]E^{\mu_0,x_0}[\tilde{c}(x_k, \mu(x_k))], \end{aligned}$$

using the independence of transition times and the system's trajectory $(x_0, u_0, x_1, \ldots)$. Finally, we obtain

$$\begin{aligned} J(\mu, x_0) &= \sum_{k=0}^{\infty}(\frac{\nu}{\alpha+\nu})^k \frac{1}{\alpha+\nu} E^{\mu,x_0}[\tilde{c}(x_k, \mu(x_k))] \\ &= E^{\mu,x_0}[\sum_{k=0}^{\infty}(\frac{\nu}{\alpha+\nu})^k \frac{1}{\alpha+\nu}\tilde{c}(x_k, \mu(x_k))] \\ &= E^{\mu,x_0}[\sum_{k=0}^{\infty} \beta^k c(x_k, \mu(x_k))]. \end{aligned}$$

We now address the question of whether a similar uniformization procedure is possible with a risk-sensitive criterion given by

$$J^\gamma(\mu, x_0) = E^{\mu,x_0}[\exp(\gamma \int_0^\infty \exp(-\alpha t)\tilde{c}(x(t), \mu(x(t)))dt)], \ \alpha > 0. \tag{20}$$

Proceeding in a similar fashion, again we set the time of the kth transition to be time k. We have once again a discount factor $\beta = \frac{\nu}{\alpha+\nu}$, and transition probability matrix $P(u)$ given by

$$P_{ij}(u) = \begin{cases} \frac{\nu_i(u)}{\nu}\tilde{P}_{ij}, & \text{if } i \neq j, \\ 1 - \frac{\nu_i(u)}{\nu}, & \text{if } i = j. \end{cases} \tag{21}$$

The criterion (6) can be re-expressed as follows:

$$\begin{aligned} J^\gamma(\mu, x_0) &= E^{\mu,x_0}[\exp(\gamma \int_0^\infty \exp(-\alpha t)\tilde{c}(x(t), \mu(x(t)))dt)] \\ &= E^{\mu,x_0}[\exp(\gamma \sum_{k=0}^{\infty} \int_{t_k}^{t_{k+1}} \exp(-\alpha t)\tilde{c}(x_k, \mu(x_k))dt)] \\ &= E^{\mu,x_0}[\exp(\gamma \sum_{k=0}^{\infty} \int_{t_k}^{t_{k+1}} \exp(-\alpha t)dt \cdot \tilde{c}(x_k, \mu(x_k)))]. \end{aligned}$$

Since $E[\int_{t_k}^{t_{k+1}} \exp(-\alpha t)dt] = (\frac{\nu}{\alpha+\nu})^k \frac{1}{\alpha+\nu}$, we have

$$J^\gamma(\mu, x_0) = E^{\mu,x_0}[\exp(\gamma \sum_{k=0}^{\infty}(\frac{\nu}{\alpha+\nu})^k A_k \frac{1}{\alpha+\nu}\tilde{c}(x_k, \mu(x_k)))], \tag{22}$$

where

$$A_k = (\frac{\alpha+\nu}{\nu})^k(\alpha+\nu)\int_{t_k}^{t_{k+1}} \exp(-\alpha t)dt. \tag{23}$$

The sequence of random variables $(A_0, A_1, \ldots)$, while correlated, is independent of the system's trajectory $(x_0, u_0, x_1, \ldots)$. Thus, the cost per stage is given by

$$c_k(i, u) = \beta^k A_k \frac{1}{\alpha+\nu}\tilde{c}(i, u), \tag{24}$$

where A_k is a unit mean random variable independent of the system trajectory.

We see that by a similar uniformation procedure, risk-sensitive uniformization transforms a continuous-time MDP formulation to a discrete-time MDP formulation. However, the result of the uniformization is slightly more complex than in the risk-neutral case, since we no longer have a cost per stage at time k that is a deterministic function of the state x_k and the control u_k.

Unfortunately, the added generality of the discrete-time MDP formulation leads to a much more difficult optimal control problem. In the risk-neutral setting, it is equivalent to replace A_k by its mean (in general, by its conditional mean given x_k and u_k). In the minimax setting, one can replace A_k by the supremum of its support, i.e. the supremum of the set of values for which there is a non-zero probability density. In the risk-sensitive setting, either simplication in general will lead to a sub-optimal controller. An effective methodology to handle the risk-sensitive optimal control problem for MDPs with a non-deterministic cost structure remains a topic for future research. It would then be possible to effectively address continuous-time queueing problems with a risk-sensitive criterion, and to try to develop risk-sensitive versions of threshold policies for routing, a generalized μC-rule for priority assignments, etc. In the example considered in the next section, we will consider a discrete-time queueing formulation.

4 An Example

We now consider a simple queueing system given in [Ber95]. For this system, we will consider both risk-neutral and risk-sensitive criteria for optimal control.

The system evolves over a finite horizon N, and there is a single queue with room for up to n customers. The probability p_m of m arrivals during a time period is known. If the queue is full, additional customers are turned away. The numbers of arrivals during separate time periods are independent. At the start of each time period, service is undertaken on a single customer (when the queue is nonempty), and is completed with probability q_f if fast service is used, and q_s if slow service is used.

At time k, the state of the system x_k is given by the number of customers in the queue. The control u_k is the type of service used during the k^{th} period. We let $u = 1$ denote fast service and $u = 2$ denote slow service. The cost per stage at time k is given by $c(x_k) + \tilde{c}(u_k)$, for $k = 0, \ldots, N-1$. There is a terminal cost $c_N(x_N)$ for the

x_N customers left in the queue at time N. The risk-neutral criterion is given by

$$J(\mu, x_0) = E^{\mu,x_0}[\sum_{k=0}^{N-1} (c(x_k) + \tilde{c}(u_k)) + c_N(x_N)], \tag{25}$$

and the risk-sensitive criterion is given by

$$J^\gamma(\mu, x_0) = E^{\mu,x_0}[\exp(\gamma[\sum_{k=0}^{N-1} (c(x_k) + \tilde{c}(u_k)) + c_N(x_N)])], \ \gamma > 0. \tag{26}$$

We will determine the optimal policy with respect to (25) and (26) for various choices of the risk-sensitive parameter γ . The computations are done with subroutines implemented in **MATLAB**. With appropriate modifications to handle the risk-sensitive criterion, we use **SSYSCODE**, a recently developed **MATLAB** software library for (risk-neutral) stochastic system control (see [FGCG96]).

We set $c(i) = i$, $\tilde{c}(1) = 1$, and $\tilde{c}(2) = 0$. That is, fast service is more expensive than slow service. Also, we set the horizon size to $N = 10$, and the terminal cost to $c_N(0) = 0$, $c_N(1) = 2$, $c_N(2) = 8$, $c_N(3) = 20$. Finally, we set $n = 3$, $p_0 = 0.2$, $p_1 = 0.5$, $p_2 = 0.3$, $p_m = 0$ for $m > 2$, $q_f = 0.8$, and $q_s = 0.3$. It can be shown that this leads to the following probability transition matrices:

$$P(1) = \begin{bmatrix} 0.20 & 0.50 & 0.30 & 0.00 \\ 0.16 & 0.44 & 0.34 & 0.06 \\ 0.00 & 0.16 & 0.44 & 0.40 \\ 0.00 & 0.00 & 0.16 & 0.84 \end{bmatrix}, \quad P(2) = \begin{bmatrix} 0.20 & 0.50 & 0.30 & 0.00 \\ 0.06 & 0.29 & 0.44 & 0.21 \\ 0.00 & 0.06 & 0.29 & 0.65 \\ 0.00 & 0.00 & 0.06 & 0.94 \end{bmatrix}.$$

The following tables summarize the risk-neutral optimal policy and the risk-sensitive optimal policy for $\gamma = 0.1$. We denote by an asterisk (*) those control actions that differ in the two policies.

State	k=0	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9
0	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	1*	1*	1*	1*	1	1	1
3	2	2	2	2	2	2	2	2	2	1*

Risk-Neutral

State	k=0	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9
0	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2*	2*	2*	2*	1	1	1
3	2	2	2	2	2	2	2	2	2	2*

Risk-Sensitive, $\gamma = 0.1$

As γ is increased further, we find that the trend of increased use of slow service continues. For $\gamma > 1.3$, we find that the optimal policy is always to use slow service. This is consistent with what we expect. Indeed, for large γ, recall that the risk-sensitive criterion approaches the minimax criterion. Since in this example the set of feasible trajectories does not depend on the choice of control, the optimal minimax policy will be always to use the slow service.

Let us now consider an infinite horizon version of the previous example. The cost per stage at time k is given by $\beta^k c(x_k) + \beta^k \tilde{c}(u_k)$, for $k = 0, 1, \ldots$. We set $c(i) = i$, $\tilde{c}(1) = 1$, $\tilde{c}(2) = 0.5$, and $\beta = 0.9$. The risk-neutral optimality criterion is given by:

$$J(\mu, x_0) = E^{\mu, x_0}[\sum_{k=0}^{\infty} \beta^k (c(x_k) + \tilde{c}(u_k))]. \tag{27}$$

Since the classical risk-sensitive criterion (2) will lead to a non-stationary optimal policy, we will compare the risk-neutral optimal policy to the Porteus risk-sensitive optimal policy. This policy satisfies the following:

$$h_k^\gamma(i) = \min_{\mu}\{c(i) + \tilde{c}(u) + \frac{\beta}{\gamma}\log[\sum_j P_{ij}(u)\exp(\gamma h_{k+1}^\gamma(j))], \; k = 0, \ldots, \tag{28}$$

for $i \in X$. We will also consider the large-risk limit of the Porteus formulation. It can be shown that this leads to the following optimality equation (see [Cor97] for details):

$$\bar{h}(i) = \min_{u \in U}\{c(i) + \tilde{c}(u) + \beta \max_{j \in \tilde{X}(i,u)} \bar{h}(j)\}, \tag{29}$$

for $i \in X$, where $\tilde{X}(i, u)$ is the set of states that the system reaches in one transition with nonzero probability, given that it is in state i and control u is used. As with the minimax criterion (4), note that in this context it is equivalent to consider a probabilistic or determininistic description of uncertainty. Indeed, the significance of action $u \in U$ in the optimality equation (29) is that it determines the set of feasible next states $\tilde{X}(i, u)$. The probability with which each state in $\tilde{X}(i, u)$ is reached is of no importance.

The following table summarizes the results obtained by our modified version of the **SSYSCODE** subroutines. The optimal policies are determined through policy iteration.

State	Risk-Neutral	Risk-Sensitive ($\gamma = 0.5$)	Minimax
0	2	2	2
1	1	1	2
2	1	2	2
3	2	2	2

Infinite Horizon

We find that for small values of γ, as expected, the Porteus risk-sensitive optimal policy satisfying (28) is the same as the risk-neutral optimal policy. As we increase γ further, this is no longer the case. For instance, for $\gamma = 0.5$ we find the optimal policy given by the middle column in the table. As γ becomes sufficiently large, as expected we find that the Porteus risk-sensitive optimal policy and the Porteus minimax optimal policy satisfying (29) are the same.

As in the finite horizon setting, for this particular example we find that an increase in the risk-sensitive parameter leads to an increased use of slow service. Also, we find that the Porteus minimax optimal policy is always to choose slow service. As in the finite horizon setting, the intuition for this is the fact that the choice of control does not alter the set of feasible trajectories for the system.

5 Concluding Remarks

The classical risk-sensitive approach to stochastic control introduces additional flexibility in the choice of an optimality criterion over the risk-neutral approach. In the MDP setting, the additional generality of the formulation does not increase the complexity of controller synthesis for finite horizon problems. See [Cor97] for details.

On the infinite horizon, in the discounted cost setting, a key difficulty with the classical risk-sensitive approach is the lack of a stationary optimal policy in general. A more general framework has been introduced in this paper, in the case where the state of the system is observed. The framework has the classical approach as a special case; other special cases correspond to previously introduced alternate approaches to risk-sensitive control [Por75], [KP78], [Eag75]. The alternate approaches lead to the existence of a stationary optimal policy, and policy and value iteration tecniques can be used for controller synthesis. The general framework in the partially observed setting is discussed further in [Cor97].

The large-risk limit of the risk-sensitive approach leads to a minimax control problem. Thus risk-sensitive control provides an important link between stochastic and robust, worst-case control approaches. Further, the risk-sensitive approach can be seen as a way to trade off the conflicting objectives of minimizing the expected cost incurred and the worst-case cost incurred.

The standard uniformization methodology, which allows one to transform a continuous-time system description into a discrete-time one, requires that the optimal control criterion be of the risk-neutral type. By generalizing the methodology, we show that it is possible to consider a risk-sensitive criterion, though the resulting discrete-time system is significantly more complex than in the risk-neutral approach. Thus, in this paper we have applied the tools of risk-sensitive control to a discrete-time queueing example.

Acknowledgement This research was partially supported by the National Science Foundation under Grant EEC 9402384.

References

[Ber95] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 1995.

[BJam] J. S. Baras and M. R. James. Robust and risk-sensitive output feedback control for finite state machines and hidden markov models. *Journal of Mathematical Systems, Estimation, and Control*, to appear.

[Cor97] S. P. Coraluppi. *Optimal Control of Markov Decision Processes for Performance and Robustness.* PhD thesis, University of Maryland, 1997.

[CS87] K. J. Chung and M. J. Sobel. Discounted mdp's: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25:49–62, 1987.

[Eag75] J. N. Eagle. *A Utility Criterion for the Markov Decision Process.* PhD thesis, Stanford University, 1975.

[FGCG96] E. Fernández-Gaucherand, J. Choi, and D. Gerhart. SSYSCODE: Stochastic system control and decision algorithms software library, fortran and MATLAB versions. Technical report, Department of Systems and Industrial Engineering, University of Arizona, 1996.

[FGMar] E. Fernández-Gaucherand and S. I. Marcus. Risk-sensitive optimal control of hidden markov models: Structural results. *IEEE Transactions on Automatic Control*, to appear.

[GD88] K. Glover and J. C. Doyle. State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity. *Systems and Control Letters*, 11:167–172, 1988.

[HM72] R. A. Howard and J. E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.

[HS84] D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research, Vol. II: Stochastic Optimization.* McGraw-Hill, 1984.

[HS95] L. P. Hansen and T. J. Sargent. Discounted linear exponential quadratic gaussian control. *IEEE Transactions on Automatic Control*, 40:968–971, 1995.

[KP78] D. M. Kreps and E. L. Porteus. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, 46(1):185–200, 1978.

[Por75] E. Porteus. On the optimality of structured policies in countable stage decision processes. *Management Science*, 22(2):148–157, 1975.

Construction of Convolutional Codes using Methods from Linear Systems Theory *

Joachim Rosenthal
Department of Mathematics,
University of Notre Dame,
Notre Dame, Indiana 46556.
e-mail: Rosenthal.1@nd.edu

Roxana Smarandache
Department of Mathematics,
University of Notre Dame,
Notre Dame, Indiana 46556.
e-mail: rsmarand@nd.edu

September 26, 1997

Abstract

Using two different matrix pencil descriptions two new classes of convolutional codes with designed lower free distance are introduced.

Keywords: Convolutional codes, linear systems, generalized first order representations, BCH codes, Alternant codes, Grothendieck quot scheme.

1 Introduction

Until recently there did exist very few constructions of convolutional codes with a designed free distance. The constructions which did exist were mainly based on some quasi cyclic constructions of block codes as e.g. in [1, 2, 4, 6] or they made use of the parity check matrix $H(z)$ of the convolutional code as e.g. in [8, 15]. A survey of some of these results can be found in the monograph of Piret [7].

In [12, 14, 16] the first author together with York and Schumacher used some basic ideas in systems theory to derive some convolutional codes of Reed Solomon and BCH type. Different from the previous work in this area convolutional codes were constructed using a first order description of the encoder. It was explained in [14] that the algebraic structure of those codes can be used to algebraically decode those codes.

The paper is structured as follows: In Section 2 we review some of the results reported in [12, 14, 16] and we give conditions on the first order descriptions leading to 'desirable'

---

*Both authors were supported in part by NSF grant DMS-96-10389.

codes both from the point of view of the designed free distance and the possibility to decode those codes. In Section 3 we show how a small adaptation in the construction presented in [12] will lead to a class of convolutional codes of alternant type. Finally in Section 4 we show that some generalized first order representations well known in the systems literature can be used to derive some convolutional codes with excellent free distance.

2 First order representations of convolutional codes having the form (A, B, C, D) and what makes them desirable

In this paper, following [12, 16], we will define a convolutional code as a submodule $\mathcal{C} \subseteq \mathbb{F}^n[z]$. Since $\mathbb{F}[z]$ is a PID and since $\mathcal{C}$ is a submodule of the free module $\mathbb{F}^n[z]$ the code $\mathcal{C}$ is free as well and it has a well defined rank k. Let $\{g_1(z), \ldots, g_k(z)\} \subset \mathbb{F}^n[z]$ be a basis of the free module $\mathcal{C}$ and let $G(z)$ be the $n \times k$ polynomial matrix whose ith column consists of the polynomial vector $g_i(z); i = 1, \ldots, k$. The convolutional code $\mathcal{C}$ is then equivalently defined as:

$$\mathcal{C} = \left\{ v(z) \in \mathbb{F}^n[z] \mid v(z) = G(z)m(z),\ m(z) \in \mathbb{F}^k[z] \right\}.$$

As usual we will call $G(z)$ an *encoder* of the convolutional code $\mathcal{C}$ and we will say that $\mathcal{C}$ has *rate* k/n if the rank of the module $\mathcal{C}$ is k. If $\tilde{G}(z)$ is a second encoder then there exists a unimodular matrix $U(z) \in Gl_k(\mathbb{F}[z])$ such that $\tilde{G}(z) = G(z)U(z)$. The *free distance* of a convolutional code is defined as

$$d_f(\mathcal{C}) := \min\{\mathrm{wt}(v(z)) \mid v(z) \in \mathcal{C}, v(z) \neq 0\},$$

where wt denotes as usual the weight of a code word. On the side of the rate and the free distance of a convolutional code there is another important parameter called the *complexity*. The complexity δ of a convolutional code $\mathcal{C}$ is defined as the largest degree δ of the $k \times k$ full size minors of one and therefore any encoder $G(z)$. We can naturally identify convolutional codes of complexity zero with block codes. We define a convolutional code to be *observable* if one and therefore every encoder $G(z)$ is right prime. If $G(z)$ is an encoder of an observable convolutional code then $G(z)$ is necessarily a non-catastrophic encoder.

A major design problem in the area of convolutional codes is the construction of an observable code having largest free distance among all possible convolutional codes of a fixed rate and a fixed complexity.

It has been pointed out in [12] that after a possible permutation of the coordinates of $v(z)$, the code $\mathcal{C}$ can be represented through a familiar looking input/state/output description. For this let:

$$v(z) = v_0 z^{\gamma} + v_1 z^{\gamma-1} + \ldots + v_{\gamma};\ v_t \in \mathbb{F}^n, t = 0, \ldots, \gamma.$$

If one partitions the vector v_t into $v_t = \binom{y_t}{u_t}$, where y_t has $n-k$ components and u_t has k components then the convolutional code is equivalently described by the familiar looking '(A,B,C,D)' representation

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t \\ y_t &= Cx_t + Du_t, \quad x_0 = 0, \ x_{\gamma+1} = 0. \end{aligned} \tag{2.1}$$

If $\mathcal{C}$ has complexity δ then it is possible to choose the matrices (A,B,C,D) of size $\delta \times \delta$, $\delta \times k$, $(n-k)\times\delta$ and $(n-k)\times k$ respectively. A representation having these sizes is called a *minimal representation* and it is algebraically characterized through the condition that (A,B) forms a *controllable* matrix pair, i.e.

$$\operatorname{rank}\begin{pmatrix} B & AB \dots & A^{\delta-1}B \end{pmatrix} = \delta. \tag{2.2}$$

As explained in [12] the convolutional code $\mathcal{C}$ defined through (2.1) represents an observable code if and only if the matrix pair (A,C) forms an *observable pair*, i.e.

$$\operatorname{rank}\begin{pmatrix} C \\ CA \\ \vdots \\ CA^{\delta-1} \end{pmatrix} = \delta. \tag{2.3}$$

The following theorem gives a general condition on matrices (A,B,C,D) which guarantees a designed distance d for a convolutional code $\mathcal{C}$:

Theorem 2.1 ([14, 16]) *Let d be a positive integer and let*

$$T := d\left\lceil \frac{\delta}{n-k} \right\rceil + 1. \tag{2.4}$$

Assume the matrices (A,B) form a controllable pair with the property that for any integer $\tau > 0$ the matrix

$$(A^{\tau}B \ \ A^{\tau+1}B \ \ \dots \ \ A^{\tau+T-1}B) \tag{2.5}$$

defines the parity check matrix of a block code of distance d. If in addition (A,C) forms an observable pair then the convolutional code defined in (2.1) *has designed free distance at least d.*

The proof of this theorem was carried out in [12] in the situation where (2.5) represents a maximum distance separable code. In the general situation the idea is the same and details are given in [14, 16].

If for every integer τ the block code defined through (2.5) comes with an efficient (algebraic) decoding algorithm then it was shown in [14] that this algebraic decoding algorithm can be used to arrive at an algebraic decoding algorithm of the corresponding convolutional code. Both from the point of view of construction as well as from the point of view of decoding it is therefore desirable to choose matrices (A,B) where the block codes defined in (2.5) have some desirable distances and decoding algorithms.

3 A class of convolutional codes of alternant type

In this section we show how to choose matrices (A, B) such that the block codes appearing in (2.5) are of alternant type. First recall from [12] a Reed-Solomon type construction. For this assume that $d = \delta$, and that α is a primitive of the field $\mathbb{F}_q$ where the number of field elements satisfies $q > Tk$. Define:

$$A := \begin{pmatrix} \alpha^k & 0 & \cdots & 0 \\ 0 & \alpha^{2k} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \alpha^{ck} \end{pmatrix}, \quad B := \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{k-1} \\ 1 & \alpha^2 & \alpha^4 & \cdots & \alpha^{2(k-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha^c & \alpha^{2c} & \cdots & \alpha^{c(k-1)} \end{pmatrix}. \tag{3.1}$$

With those choices it is guaranteed that for each $\tau > 0$ the matrices appearing in (2.5) are of Reed-Solomon type, in particular the parity check matrices appearing in (2.5) describe MDS codes.

We can now adapt this construction slightly. Let h be any nonzero field element and define:

$$A := \begin{pmatrix} h^k\alpha^k & 0 & \cdots & 0 \\ 0 & h^k\alpha^{2k} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & h^k\alpha^{ck} \end{pmatrix}, \quad B := \begin{pmatrix} 1 & h\alpha & h^2\alpha^2 & \cdots & h^{k-1}\alpha^{k-1} \\ 1 & h\alpha^2 & h^2\alpha^4 & \cdots & h^{k-1}\alpha^{2(k-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & h\alpha^c & h^2\alpha^{2c} & \cdots & h^{k-1}\alpha^{c(k-1)} \end{pmatrix}. \tag{3.2}$$

One immediately verifies that the block codes appearing in this case in (2.5) have now an 'alternant structure'. One can show that the subfield construction carried through in the situation when (A, B) are of the form (3.1) still can be carried through in this more general situation and the decoding algorithm presented in [14] still applies.

The distinct advantage to work with general matrices of type (3.2) is the fact that BCH block codes are known to be asymptotically bad whereas it is known that there are Alternant block codes which asymptotically are good.

We will address those issues in future work.

4 Construction of convolutional codes using generalized first order descriptions

In this section we explain that it is possible to construct some good convolutional codes using a purely module theoretic approach. In particular we will in this section not use an (A, B, C, D) description for the convolutional code.

It was first pointed out in [13] that a submodule $\mathcal{C} \subset \mathbb{F}^n[z]$ having rank k and complexity δ can also be represented in the following way:

Theorem 4.1 *Assume $\mathcal{C} \subseteq \mathbb{F}^n[z]$ is a rate $\frac{k}{n}$ convolutional code of complexity δ. Then there exist matrices P, Q having sizes $\delta \times (\delta + k)$, and a matrix R having size $n \times (\delta + k)$ (all defined over $\mathbb{F}$) such that the code $\mathcal{C}$ is described by*

$$\mathcal{C} = \{v(z) \in \mathbb{F}^n[z] \mid v(z) = R\zeta(z),\ (zP + Q)\zeta(z) = 0\}. \tag{4.1}$$

Moreover the following conditions are satisfied:

(1) P has full row rank;

(2) $\begin{bmatrix} P \\ R \end{bmatrix}$ has full column rank;

(3) $\begin{bmatrix} zP + Q \\ R \end{bmatrix}$ is right prime.

The representation given in Theorem 4.1 is dual to the '(K, L, M)' representation given in [12, Theorem 3.1.]. Both the (K, L, M) representation and above representation are well studied in the systems theory literature and we refer e.g. to [3], where also further references can be found.

Clearly if the triple of matrices (P, Q, R) have the sizes of Theorem 4.1 then (4.1) defines a submodule of $\mathbb{F}^n[z]$, i.e. a convolutional code. If (P, Q, R) is not a minimal triple then this convolutional code has in general neither rank k nor complexity δ. The next theorem will show that a convolutional code defined by (4.1) has rank k and has complexity δ as soon as the minimality conditions of Theorem 4.1 are satisfied.

It arises the question of how 'unique' the representation (4.1) is. The following Theorem answers this question. The theorem is well known in the systems literature (see e.g. [3, Theorem 4.34]) and it is the dual version of a theorem given in [12]. In systems theory the base field $\mathbb{F}$ is usually assumed to consist of the real numbers however a study of the proof in [3] verifies that the result is valid over any base field:

Theorem 4.2 *The matrices (P, Q, R) introduced in Theorem 4.1 are unique in the following way: if $(\tilde{P}, \tilde{Q}, \tilde{R})$ is a second triple of matrices describing the code $\mathcal{C}$ through*

$$\mathcal{C} = \{v(z) \in \mathbb{F}^n[z] \mid v(z) = \tilde{R}\zeta(z),\ (z\tilde{P} + \tilde{Q})\zeta(z) = 0\}.$$

and if (P, Q, R) and $(\tilde{P}, \tilde{Q}, \tilde{R})$ both satisfy the minimality conditions of Theorem 4.1 then there exist unique invertible matrices T and S such that

$$(\tilde{P}, \tilde{Q}, \tilde{R}) = (SPT^{-1}, SQT^{-1}, RT^{-1}). \tag{4.2}$$

Moreover every minimal triple which describes the convolutional code $\mathcal{C}$ is of the form (4.2). Finally the triple (P, Q, R) (and hence any triple of the form (4.2)) describes a rate k/n convolutional code of complexity δ.

For the purpose of constructing convolutional codes of a certain rate and a certain complexity it is therefore no limitation to work with minimal triples (P, Q, R) satisfying the minimality conditions of Theorem 4.1. Algorithms are available to go from this description to a polynomial encoder description.

There remains of course the problem of identifying those convolutional codes which are observable, i.e. in particular those codes which have non-catastrophic encoders. The following Lemma provides an answer to this question:

Lemma 4.3 *A matrix triple (P, Q, R) satisfying the minimality conditions of Theorem 4.1 defines an observable convolutional code if and only if the matrix pencil $[zP + Q]$ is left prime.*

Proof: Direct consequence of [11, Theorem 3.3]. □

In [14] it was shown that the Reed Solomon type convolutional codes achieve a free distance which is approximately $\frac{k}{n}$ times the best possible free distance found among all convolutional codes of rate $\frac{k}{n}$ and complexity δ. For low rate those code are therefore not as good. Justesen did construct in [2] rate $\frac{1}{n}$ convolutional codes of degree δ with optimal free distance. In the remainder of this section we will show how to construct (using the representations of Theorem 4.1) some low rate convolutional codes having a free distance superior to the free distance of the codes constructed in [12]. We start with a Lemma:

Lemma 4.4 *There exist matrices (P, Q) having sizes $\delta \times (\delta + k)$ such that the following is true:*

1. *If $\zeta(z) \in \mathbb{F}^{\delta+k}$ is a polynomial vector whose degree is strictly less than $\lfloor \frac{\delta}{k} \rfloor$ and if $(zP + Q)\zeta(z) = 0$ then necessarily $\zeta(z) = 0$.*
2. *$[zP + Q]$ is left prime.*

There are many ways of constructing such matrices (P, Q). A particular example could be $P = (I_\delta, 0)$ and $Q = (A, B)$ where (A, B) are either defined by (3.1) or by (3.2).

Now we can show that for certain low rates we can obtain better distance codes than the ones presented in [12, 14]:

Theorem 4.5 *Let (P, Q) be defined as in Lemma 4.4. Assume $n > \delta + k$ and let R be a matrix of size $n \times (\delta + k)$. If the columns of R form the generator matrix of a maximum distance separable code then the code defined by the triple (P, Q, R) has rate $\frac{k}{n}$, complexity δ and free distance at least $\left(\lfloor \frac{\delta}{k} \rfloor + 1\right)(n - \delta - k + 1)$.*

Proof: Let $v = v_0 + v_1 z + \cdots + v_\gamma z^\gamma$ be a code vector of degree γ and assume that $v_0 \neq 0$. Then, using the sliding matrix representation, we have

$$\begin{bmatrix} v_o \\ v_1 \\ \vdots \\ v_\gamma \end{bmatrix} = \begin{bmatrix} R & 0 & \cdots & 0 \\ 0 & R & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & R \end{bmatrix} \begin{bmatrix} \zeta_0 \\ \zeta_1 \\ \vdots \\ \zeta_\gamma \end{bmatrix}.$$

The first $\left(\lfloor\frac{\delta}{k}\rfloor + 1\right)$ components of the vector $\zeta(z)$ have to be nonzero by our assumption on (P, Q). Because the columns of R define a MDS code the claim follows. $\square$

We conclude this section with some remarks about the algebraic geometric ingredient of the pencil descriptions considered in this section.

As it was explained in [10] a submodule of rank k and complexity δ in $\mathbb{F}^n[z]$ describes a quotient sheaf of rank k and degree δ. By a general theorem of Grothendieck it is possible to equip the set of all rank k submodules (quotient sheafs) of complexity (i.e. degree) δ with the structure of a scheme. Such a scheme is referred to as a *quot scheme* in the algebraic geometry literature. The quot scheme which parameterizes the rank k submodules of complexity δ turns out to be a smooth projective variety [9].

If the degree $\delta = 0$ the Grothendieck quot scheme is exactly the Grassmann variety $\text{Grass}(k, \mathbb{F}^n)$ consisting of all k dimensional subspaces of the vector space $\mathbb{F}^n$. This variety parameterizes the set of all linear block codes of rate $\frac{k}{n}$ defined over the field $\mathbb{F}$. For an arbitrary complexity δ the Grothendieck quot scheme parameterizes in a natural way all rate $\frac{k}{n}$ convolutional codes of complexity δ.

Linear systems described by matrix triples (P, Q, R) have been studied widely in the systems literature and probably the most comprehensive account is given in the monograph of Kuijper [3]. It was pointed out by Lomadze [5] that a matrix pencil of the form $\left[\begin{smallmatrix} zP+Q \\ R \end{smallmatrix}\right]$ represents exactly the linear free resolution of the associated quotient sheaf and in this way such matrix pencils appear naturally in the algebraic geometry literature as well. Finally we would like to note that we can view (4.2) as a group action of the reductive group $Gl_\delta \times Gl_{\delta+k}$ on the vector space consisting of all matrix triples (P, Q, R) of a fixed size. The uniqueness Theorem 4.2 expresses the fact that the group orbits in (4.2) correspond to the submodules of $\mathbb{F}^n[z]$, i.e. the convolutional codes.

Actually much more is true: The geometric quotient in the sense of GIT (=geometric invariant theory) induced by the group action (4.2) is exactly the Grothendick quot scheme. The minimality conditions provided in Theorem 4.1 guarantee that the associated orbit is a 'stable orbit' in the sense of GIT. This is true for an arbitrary base field and this statement is a geometric formulation of the uniqueness Theorem 4.2. The reader who is interested in more details is referred to [10].

References

[1] J. Justesen. New convolutional code constructions and a class of asymptotically good time-varying codes. *IEEE Trans. Inform. Theory*, IT-19(2):220–225, 1973.

[2] J. Justesen. An algebraic construction of rate $1/\nu$ convolutional codes. *IEEE Trans. Inform. Theory*, IT-21(1):577–580, 1975.

[3] M. Kuijper. *First-Order Representations of Linear Systems.* Birkhäuser, Boston, 1994.

[4] Y. Levy and D.J. Costello Jr. An algebraic approach to constructing convolutional codes from quasi-cyclic codes. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 14:189–198, 1993.

[5] V. Lomadze. Finite-dimensional time-invariant linear dynamical systems: Algebraic theory. *Acta Appl. Math*, 19:149–201, 1990.

[6] J. L. Massey, D.J. Costello, and J. Justesen. Polynomial weights and code constructions. *IEEE Trans. Inform. Theory*, IT-19(1):101–110, 1973.

[7] Ph. Piret. *Convolutional Codes, an Algebraic Approach.* MIT Press, Cambridge, MA, 1988.

[8] Ph. Piret. A convolutional equivalent to Reed-Solomon codes. *Philips J. Res.*, 43(3-4):441–458, 1988.

[9] M. S. Ravi and J. Rosenthal. A smooth compactification of the space of transfer functions with fixed McMillan degree. *Acta Appl. Math*, 34:329–352, 1994.

[10] M. S. Ravi and J. Rosenthal. A general realization theory for higher order linear differential equations. *Systems & Control Letters*, 25(5):351–360, 1995.

[11] M. S. Ravi, J. Rosenthal, and J. M. Schumacher. Homogeneous behaviors. *Math. Contr., Sign., and Syst.*, 10:61–75, 1997.

[12] J. Rosenthal, J. M. Schumacher, and E.V. York. On behaviors and convolutional codes. *IEEE Trans. Inform. Theory*, 42(6):1881–1891, 1996.

[13] J. Rosenthal and E.V. York. Linear systems defined over a finite field, dynamic programming and convolutional codes. In *Proc. of the IFAC Conference on System Structure and Control*, pages 466–471, Nantes, France, July 1995.

[14] J. Rosenthal and E.V. York. BCH convolutional codes. Preprint, October 1997.

[15] R.M. Tanner. Convolutional codes from quasi-cyclic codes: A link between the theories of block and convolutional codes. Computer Research Laboratory, Technical Report, USC-CRL-87-21, November 1987.

[16] E.V. York. *Algebraic Description and Construction of Error Correcting Codes, a Systems Theory Point of View.* PhD thesis, University of Notre Dame, 1997. Available at http://www.nd.edu/~rosen/preprints.html.

Minimum Weights and Correctable Error Patterns of Convolutional Codes

Jørn Justesen
Department of Telecommunication
Building 371 Technical University of Denmark, DK-2800 Lyngby, Denmark

Abstract: We discuss some properties of the minimum distances of convolutional codes and ways of relating sets of correctable error pattern to these distances. In particular we modify the Hamming distance between sequences to make a metric space of infinitely long strings and consider sets of error patterns of bounded distance in this space.

I. Minimum distance decoding of convolutional codes.

A convolutional encoder is described by a linear encoding rule of the type

$$y_t = \sum x_{t-j} \; g_j$$

where y are length n output blocks, x are length k input blocks, and g are generator matrices.

The minimum (row) distances, d_j^r, are the minimum Hamming weights of nonzero code-words such that block y_1 is the first nonzero block, and the encoder returns to the zero state after block y_j.

We shall refer to minimum distance decoding [1] of a convolutional code as a decoding that corrects any error patterns with the property that any segment of j consecutive blocks contains less than $d_j^r/2$ errors. If the number of errors violates this rule, there is some particular location of the errors that would cause a decoding error or an ambiguous decision.

For rates less than R_0, the performance of convolutional codes is determined by the code-words of weight d_{free}, and most error events occur at a critical length where such words exist [2]. Contrary to what is commonly stated, the events are not always very short. In comparison, minimum weight error patterns are important in block codes at much lower rates.

II. Minimum row distances of convolutional codes.

As discussed in [1], the minimum (extended) row distances of a convolutional codes may be expressed as

$$d_j^r = Aj + B + C(j) + D(j)$$

where A and B are rational constants, D(j) is zero for j>M for some finite M, and C(j) is a periodic function with period L. As pointed out by one of the reviewers of [1], the claim that C is periodic does not seem to be proved. We shall give a proof based on calculating the sequence by means of a modification of the Viterbi algorithm. In this way we also find some properties of the periodic function.

Let the linear encoder for the convolutional code have states s_0 , s_1 , s_2 , ... , where in particular s_0 is the all zero state. For each depth in the decoding trellis, n>0 we calculate minimum weights for the paths from state s_0 to s_j , $u_j(n)$ by initially setting $u_0(0) = 0$, removing all transitions out of s_0 for n>0, and calculating

$$u_j(n+1) = \min_i \{ u_i(n) + m_{ij} \}$$

where m_{ij} is the weight of the output block for the state transition from state s_i to s_j. It follows from the properties of the Viterbi algorithm that the row distances are obtained as $u_0(n)$.

It is well known that the constant A is found as the minimum average weight over all cycles in the state graph. Clearly there may be several cycles with the same average weight, and in that case the denominator of A is a common factor of the lengths. Since no state can appear more than once in the shortest cycle, the denominator cannot exceed the number of states.

If in each depth of the trellis the $u_j(n)$ are replaced by the normalized values

$$u_j'(n) = u_j(n) - An$$

the results are bounded. This follows from the observation that each state can be reached from any other state in a finite number of transitions.

The vector ($u_1'(n)$, $u_2'(n)$, ...$u_i'(n)$) takes values in a finite set and is followed by a unique vector at depth n+1. Consequently the sequence of these vectors is eventually periodic, and thus the sequence $u_0'(n)$ are periodic, which proves the claim about C.

The properties of the algorithm indicate that the period, L could be quite large, and in particular it is not bounded by the number of states in the encoder.

III. Distances of fixed codes.

The well known short convolutional codes with good distances have properties that are somewhat different from what we have described for typical long codes. If we neglect the periodic term and write the distances as

$$d_j^r = Aj + B$$

for sufficiently large j, we can often associate the term Aj with a short periodic sequence, and the extra weight B comes from the initial and final segments connecting the cycle to the zero state. The lower bounds [3] are derived using time varying codes where there is no periodic segment, and the weight is uniformly distributed over the length of the codeword.

Within the same approximation, the row distances of long codes are approximately Aj and the segment distances Aj-B. For short codes, column distances are typically found as the initial segment of the minimum weight sequence, and are often approximately Aj+B/2. For longer codes the term B/2 would be compensated by a low weight tail between the minimum average weight cycle and an arbitrary state.

Whereas the distribution of the weight within a minimum weight codeword is not important for maximum likely decoding, it is important in other algorithms that the row distances are high. Unfortunately the properties of fixed codes cannot be obtained by random coding arguments, and in particular it is not clear what the length of the minimum average weight cycle can be expected to be.

VI. Metric spaces.

One of the difficulties in defining minimum distance decoding is that a random sequence of errors will violate the weight condition sooner or later, and there is no simple way of bounding the length of the error event. Thus we cannot in general conclude that a particular received segment will be correctly decoded unless we make unrealistic assumptions about the distribution of errors.

Even though the weight sequences are described in a simple way, there has been little progress in using these concepts in the analysis of convolutional codes. One obstacle is the lack of algebraic structure in sets that include finite sequences of different lengths. This problem is known in the theory of sequential decoding, where the Fano metric was introduced as a way of comparing weights of finite sequences.

We shall consider the possible advantages of modifying the usual (Hamming) weight of finite sequences

$$w_h(y) = \sum w_h(y_j)$$

by introducing a sequence of decreasing, positive scaling factors, b_j

$$w(y) = \sum w_h(y_j)\, b_j$$

For simplicity we shall just refer to w as the weight of the sequence.

The scaling factors serve two purposes:

1) They reflect the decreasing importance of distant parts of the sequences on decoding the block presently being considered. We shall study the possibilities of selecting the factors in a way that accurately reflect the importance of each block on the decoding decision.

2) They may be chosen to give finite weights to all semi-infinite sequences. In this way the space of such sequences becomes a metric space [4] with distance function

$$D(x,y) = w(x+y)$$

(addition always being understood as mod 2).

Finite sequences have weights, but we cannot talk about distances between sequences of different lengths. However, due to the limited contribution from the extensions to semi-infinite sequences, we can easily obtain bounds on such distances, and this may be used in a modified version of sequential decoding.

V. Minimum distance decoding in a metric space.

In order to indicate how we would like the weight function to describe the error-correcting properties of convolutional codes we consider a simple example.

Example:

We shall consider the well known (8,4) partial unit memory code [5], but actually it is sufficient to know the sequence of row distances:

$$d_j' = 8,\ 8,\ 10,\ 12,\ 14,\ \ldots,\ (2j+4), \ldots$$

and to know that the minimum weight words have Hamming weights 4 in the first and last nonzero block and 2 in any intermediate block. A minimum weight word may be extended to a longer minimum weight word by reducing the weight of block j by two while adding a new nonzero block j+1 of weight 4. If we let $b_1 = b_2 = 1$, and $b_j = 2^{2-j}$ for j>2, then all minimum weight words have the same weight, 8. Thus as a first significant advantage we have replaced the sequence of minimum weights by a single number. If we measure weights of error patterns in the same way, it follows immediately that any error pattern of weight less than 4 is correctly decoded in the first block (the minimum distance refers only to sequences that differ in the first block). An error pattern which satisfies this condition and has the errors located as close to the beginning of the sequence as possible, has Hamming weights

$$w_h(j) = 3,\ 0,\ 1,\ 1,\ 1,\ \ldots, 0,\ \ldots$$

Clearly this is in fact the maximal weights of error patterns that the code can correct. If we consider a Hamming weight sequence like

$$w_h(j) = 2,\ 0,\ 1,\ 2,\ 2,\ 0,\ \ldots, 0,\ \ldots$$

the weight is $w = 3.25 < 4$. If we consider block 1 to be correctly decoded, and shift the error pattern left, the weight becomes w=2.5, and block 2 is also correctly decoded. However a further shift gives w=4, which indicates that the decoding may be ambiguous. This happens because there are codewords of length 3 which have Hamming weight 10, and a sequence with 5 errors can have distance 5 to at least two codewords. Thus the minimum weight accurately describes the correctable error patterns, even when sequences of arbitrary length are considered. Furthermore, even in case of an error patterns that cannot be correctly decoded, the weight may be used to indicate that certain segments are correctly decoded.

The example indicates that at least in some cases, it is possible to express the distance of a convolutional code by a single number d, and to describe the error patterns that are correctly decoded in block 1 by the condition that they have weight $w<d/2$. Compared to the condition of minimum distance decoding referred to in Section I, we have the advantage that the new condition can be used even when the Hamming weight of a segment of the error pattern is too high.

However. we cannot fit maximum likelihood decoding of long error patterns into this approach. It was proved in [1] that the largest difference between the metrics on two paths which diverge in the first block and have not merged at depth j is approximately the difference between the row and column distances. If one path has the maximal number of errors, (Aj+B)/2, there may be an alternative path with (Aj-B)/2 errors. Since the maximum likelihood decision between these two paths can be delayed arbitrarily, the metric space decoder will be forced to make a decision which may be different from that of the maximum likelihood decoder. However, as noter in Section I, the performance is determined by the decoding of short error patterns, and the degradation due to suboptimal decisions may be acceptable.

References:

1. J. Justesen, "Bounded distance decoding of unit memory codes", IEEE Trans. Info. Th., IT-41, pp. 1616-27, September 1993.

2. J. Justesen and J.D. Andersen, "Critical Lengths of Error Events in Convolutional Codes", IEEE Trans. Info. Th.

3. C. Thommesen and J. Justesen, "Bounds on Distances and Error Exponents of Unit Memory Codes", IEEE Trans. Info. Th., IT-29, pp. 637-649, September, 1983.

4. D. Lind and B. Marcus, Symbolic Dynamics and Coding, Cambridge University Press, 1995.

5. L.N. Lee,"Short unit-memory byte oriented binary convolutional codes having maximal free distance", IEEE Trans. Info. Th., IT-22, pp. 349-352, May 1976.

Multilevel Coding: Use of Hard Decisions in Multistage Decoding

Udo Wachsmann, Robert Fischer, Johannes Huber

Lehrstuhl für Nachrichtentechnik II: Digital Transmission and Mobile Communications
Universität Erlangen–Nürnberg
Cauerstrasse 7, D-91058 Erlangen, Germany
Phone: +49-9131-857667, Fax: +49-9131-858849
E-mail: wachsman@nt.e-technik.uni-erlangen.de

Abstract — **Multilevel coding (MLC) schemes with powerful softly decodable component codes like turbo codes allow very power and bandwidth efficient digital communication [WH95, FHW96, WFH97a]. But the use of the non–quantized channel output in each stage of the multistage decoding procedure leads to highly complex receivers. In this paper, the employment of hard decision decoding at several coding levels is proposed as a very efficient way to save complexity. It turns out that, for Ungerböck's labeling of signal points, the performance loss compared to soft decision decoding (in terms of capacity of the equivalent binary channel as well as in simulations) is only about 0.13 dB for ASK transmission over the AWGN channel, when hard decision decoding is employed at all but the lowest level. If error–correction techniques are exclusively employed, an MLC approach with ℓ individual binary error–correcting codes is recommended.**

1 Introduction

Multilevel coding (MLC) together with the relatively low complex multistage decoding (MSD) is well–known to be a theoretically optimum approach to coded modulation, cf. [WFH97a, WH95, HW94]. Key point is the proper assignment of rates to the individual component codes. In [WFH97a, WFH97c, FHW96, WH95] it is shown that properly designed MLC schemes with powerful softly decodable component codes lead to very power and bandwidth efficient digital communication schemes close to limits from information theory. The drawback is the relatively high complexity of the receiver due to soft decision maximum likelihood (ML) decoding in each stage. Complexity and decoding delay can be substantially reduced by using simple FEC coding at some levels instead. But hard decision decoding usually causes a substantial performance loss when compared to soft decision decoding. For example, the loss for binary antipodal signaling is about 2 dB due to hard decision decoding and an additional 1 dB due to bounded minimum distance (BMD) instead of ML decoding. In previous work, MLC schemes are mostly treated, where all component codes are either exclusively error–correcting codes like algebraic block codes, cf. e.g. [PT89] or softly decodable codes like convolutional or turbo codes [KZS94, WH95]. In this paper, a performance analysis of MLC schemes using hard decisions at some levels in MSD is presented, cf. also [WFH97b].

In Section 2 the underlying system model and the concept of equivalent channels at the individual levels of a MLC scheme is introduced. In Section 3 the capacity of coded modulation schemes with a combination of hard and soft decoded levels is derived. For the example of a three–level coded 8ASK modulation, the capacity results are discussed in Section 4. In order to confirm the theoretical results, simulation results for selected schemes are presented in Section 5. Section 6 concludes this paper.

2 System Model

Digital transmission by M–ary pulse amplitude modulation, $M = 2^\ell$, $\ell > 1$, with ℓ–level coding over the AWGN channel is considered. A binary address vector or label $\mathbf{x} = (x^0 x^1 \dots x^{\ell-1})$ is assigned to each signal point a_m out of the D–dimensional signal set $\mathbf{A} = \{a_0, \dots, a_{M-1}\}$. This mapping is usually derived from successive partitioning of the set $\mathbf{A}$ into subsets [Ung82, IH77]. Each subset at partitioning level i is uniquely labeled by the path $(x^0 \dots x^{i-1})$ in the partitioning tree from the root to the subset:

$$\mathbf{A}(x^0 \dots x^{i-1}) = \{a | a \leftrightarrow \mathbf{x} = (x^0 \dots x^{i-1} b^i \dots b^{\ell-1}), (b^i \dots b^{\ell-1}) \in \{0,1\}^{\ell-i}\}. \qquad (1)$$

Here, the discussion is restricted to schemes with regular partitioning, where the subsets at one partitioning level differ only by translation and rotation, i.e. they provide equal capacities.

The core of the MLC approach is to partition a block of data symbols into ℓ levels and to encode them individually by component codes of rates R^i and equal lengths N into the binary address symbols x^i [IH77]. The total rate of the scheme is $R = \sum_{i=0}^{\ell-1} R^i$. Key point to create powerful MLC schemes is the proper assignment of code rates R^i to the component codes. Since overall maximum likelihood decoding is not feasible in practice, the component codes are successively decoded starting with the level 0 code. Thereby, not only the channel output $y \in \mathbb{R}^D$ but also decisions $\hat{x}^j$, $j = 0, \dots, i-1$, from previous decoding stages j, $j < i$, are used for decoding at level i.

Due to the individual encoding at each level, the transmission of address vector $\mathbf{x}$ with binary digits x^i, $i = 0, \dots, \ell - 1$, over the physical channel can be virtually separated into parallel transmission of each digit x^i over ℓ equivalent channels. When hard decision decoding is used at decoding stage i, the continuous channel output y is quantized to a binary symbol $\tilde{x}^i$.

For example, let us consider the equivalent channel for transmission of symbol x^1 for an 8ASK scheme when hard decision decoding is used, see Fig. 1. With the restriction to equiprobable signal points, the optimum detection thresholds are half–way between two adjacent points. Due to the multiple representation of one binary symbol by signal points, the decision region of binary symbol $\tilde{x}^i$ is the union of the corresponding regions $\mathcal{R}_{\mathbf{A}(x^0 \dots x^{i-1})}(a)$, see Fig. 1:

$$\mathcal{R}(\tilde{x}^i | x^0 \dots x^{i-1}) = \bigcup_{a \in \mathbf{A}(x^0 \dots x^{i-1} \tilde{x}^i)} \mathcal{R}_{\mathbf{A}(x^0 \dots x^{i-1})}(a), \qquad (2)$$

where the decision region $\mathcal{R}_{\mathbf{A}}(a)$ of signal point $a \in \mathbf{A}$ is defined by

$$\mathcal{R}_{\mathbf{A}}(a) = \{y | d_E(y, a) \leq d_E(y, a'), a' \in \mathbf{A}, a \neq a'\}. \qquad (3)$$

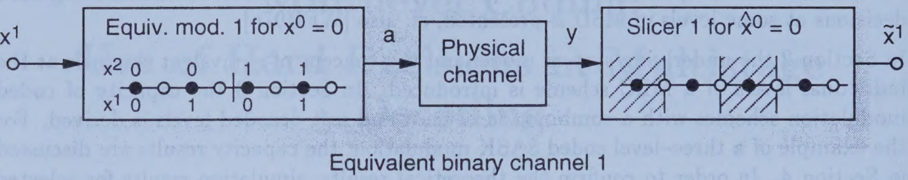


Figure 1: Equivalent BSC 1 for the transmission of symbol x^1 for an 8ASK scheme when hard decision decoding is used and symbol $x^0 = 0$ is assumed at the lowest level.

$d_E(y, a)$ denotes the Euclidean distance between y and a.

For equiprobable signal points and regular partitioning, the resulting equivalent channel is a *binary symmetric* channel (BSC). To distinguish between equivalent channel i with hard and soft decision decoding, we refer to the former one as equivalent BSC i. Capacity of the equivalent channel i without quantization is denoted by C^i, whereas capacity of the equivalent BSC i is denoted by C_H^i (index H indicates hard decision decoding).

3 Capacity

Since we consider capacities, the digits at lower levels can be assumed to be estimated without errors, as long as rates R^j are chosen not to be greater than capacities C^j or C_H^j, $j < i$, resp. Hence, the capacity of the equivalent BSC i is simply given by

$$C_H^i = 1 - \mathrm{H}_2(\varepsilon_i), \tag{4}$$

where ε_i denotes the bit error probability of the equivalent BSC i and $\mathrm{H}_2(\cdot)$ is the binary entropy function. The only task for calculating C_H^i is to determine the transition probability $\Pr\{\tilde{x}^i \neq x^i | x^i, x^0 \dots x^{i-1}\} = \varepsilon_i$. The particular choice of lower level digits $(x^0 \dots x^{i-1})$ has no influence on the transition probability since regular partitioning and equiprobable signal points are considered.

First, let us assume that $x^i = 0$ is transmitted and represented by a special signal point a out of the subset $\mathbf{A}(x^0 \dots x^{i-1}0)$. The error event of detecting $\tilde{x}^i = 1$ occurs if the channel output y falls in the decision region $\mathcal{R}_{\mathbf{A}(x^0 \dots x^{i-1})}(\tilde{a})$ of any signal point $\tilde{a}$ out of the subset $\mathbf{A}(x^0 \dots x^{i-1}1)$ representing the binary symbol $x^i = 1$. Thus, the probability of detection error is given by

$$\begin{aligned}\Pr\{\tilde{x}^i = 1 | a, x^0 \dots x^{i-1}\} &= \Pr\{\tilde{a} \in \mathbf{A}(x^0 \dots x^{i-1}1) | a, x^0 \dots x^{i-1}\} \\ &= \sum_{\tilde{a} \in \mathbf{A}(x^0 \dots x^{i-1}1)} \Pr\{\tilde{a} | a, x^0 \dots x^{i-1}\}.\end{aligned} \tag{5}$$

Here, $\Pr\{\tilde{a} | a, x^0 \dots x^{i-1}\}$ denotes the probability that y falls into $\mathcal{R}_{\mathbf{A}(x^0 \dots x^{i-1})}(\tilde{a})$, when signal point a is transmitted. Detecting any signal point $\tilde{a} \in \mathbf{A}(x^0 \dots x^{i-1}0)$, $\tilde{a} \neq a$, which also represents the symbol $x^i = 0$, does not lead to an error. The probability of detection error is upper bounded by the conditioned symbol error probability of the underlying signal subset for transmitting $a \in \mathbf{A}(x^0 \dots x^{i-1}0)$, namely

$$\begin{aligned}\Pr\{\tilde{x}^i = 1 | a, x^0 \dots x^{i-1}\} &\le \Pr\{\tilde{a} \in \mathbf{A}(x^0 \dots x^{i-1}), \tilde{a} \neq a | a, x^0 \dots x^{i-1}\} \\ &= \sum_{\tilde{a} \in \mathbf{A}(x^0 \dots x^{i-1}), \tilde{a} \neq a} \Pr\{\tilde{a} | a, x^0 \dots x^{i-1}\}. \end{aligned} \tag{6}$$

For Ungerböck's set partitioning, the most relevant case in practice, all signal points closest to the transmitted one represent the inverse binary symbol. Here, the detection error probability is mainly determined by the nearest neighbor signal points, and, hence, approximation by the symbol error probability (6) is quite tight. For example, the bit error probabilities ε_i, $i = 0, \dots, \ell - 1$, for a 2^ℓ–ary ASK scheme with $a \in \mathbf{A} = \{\pm 1, \dots, \pm(2^\ell - 1)\}$ are simply approximated by the well–known formulae for the symbol error probabilities for $2^{\ell - i}$–ary ASK.

With Eq. (5) the probability of detecting symbol $\tilde{x}^i = 1$ for transmission of symbol $x^i = 0$ is given by the expected value of $\Pr\{\tilde{x}^i | a, x^0 \dots x^{i-1}\}$ over all possible signal points a representing symbol $x^i = 0$:

$$\Pr\{\tilde{x}^i = 1 | x^i = 0, x^0 \dots x^{i-1}\} = \mathrm{E}_{a \in \mathbf{A}(x^0 \dots x^{i-1}0)} \left\{ \Pr\{\tilde{x}^i | a, x^0 \dots x^{i-1}\} \right\}. \tag{7}$$

The results derived above can simply be generalized to the case of arbitrary probability distribution of signal points and arbitrary labeling, cf. [WFH97b].

The capacity of a coded modulation scheme with a mix of hard and soft decision decoding at the individual levels is obtained by the sum over the capacities of the corresponding equivalent channels i, $i = 0, \dots, \ell - 1$. For example, consider a 3–level scheme with soft decision decoding at level 0 and hard decision decoding at levels 1 and 2:

$$C_{SHH} = C^0 + C_H^1 + C_H^2. \tag{8}$$

The index of the capacity C_{SHH} denotes the decoding method (soft or hard) at the individual levels. If the index is omitted soft decision decoding at all levels is assumed.

It was shown in [HW94, WH95, WFH97a] that, in the case of soft decision decoding at all levels, MLC together with MSD is an asymptotically optimum coding approach. However, when hard decision decoding at several levels is used, this statement is no longer valid, and, hence, the MLC approach is not necessarily optimum. Thus, the MLC/MSD transmission scheme with hard decision decoding at all levels operating on the AWGN channel is compared to a coded modulation scheme operating on a M–ary *discrete memoryless channel* (DMC). The transition probability $\Pr\{a_k | a_j\}$ for receiving symbol a_k is given by the probability that the output y of the underlying AWGN channel falls into the decision region $\mathcal{R}_{\mathbf{A}}(a_k)$ if symbol a_j is transmitted. The capacity of this scheme is given by

$$C_{DMC} = \mathrm{E}_{(a_k \in \mathbf{A}, a_j \in \mathbf{A})} \left\{ \log_2 \frac{\Pr\{a_k | a_j\}}{\Pr\{a_k\}} \right\}. \tag{9}$$

Finally, we compare all these methods of coded modulation with a traditional FEC approach where the 2^ℓ–ary modulation scheme is actually ignored. Hard decision decoding together with short bit interleaving provides a memoryless BSC. In order to minimize the bit error probability ε of the resulting BSC, Gray mapping, $a = \mathcal{M}_G(x^0, \dots, x^{\ell-1})$ is applied. The capacity reads

$$C_{FEC} = \ell \cdot (1 - \mathrm{H}_2(\varepsilon)). \tag{10}$$

For the bit error probability ε we obtain

$$\varepsilon = \frac{1}{\ell}\,\mathrm{E}_{(a_k \in \mathbf{A}, a_j \in \mathbf{A})}\left\{\Pr\{a_j\}\Pr\{a_k|a_j\} \cdot d_H(\mathcal{M}_G^{-1}(a_k), \mathcal{M}_G^{-1}(a_j))\right\}. \tag{11}$$

where $d_H(\cdot)$ denotes the Hamming distance. Such a scheme corresponds to bit interleaved coded modulation (BICM), see [CTB96], with hard decision decoding (BICMH).

4 Examples and Discussion

For example, 3–level coded 8ASK modulation with natural labeling is investigated. In Fig. 2 the

- capacity C, i.e. soft decision decoding at all levels (case SSS, reference),
- capacity C_{SHH}, i.e. soft decision decoding at level 0 and hard decision decoding at levels 1 and 2 (case SHH),
- capacity C_{HHH}, i.e. hard decision decoding at all levels (case HHH),
- capacity C_{DMC} of the 8–ary DMC (dashed line)
- capacity C_{FEC} (dashed–dotted line)

together with the corresponding capacities of the equivalent channels and the equivalent BSCs, respectively, are depicted.

Firstly, we look at the capacities C^i of the equivalent channels i and C_H^i, $i = 0, 1, 2$, of the equivalent BSCs i, respectively. At level 2, the underlying signal set for the transmission of symbol x^2 is the ordinary BPSK signal set, i.e. symbol x^2 is represented uniquely. In this case, the well–known gap of about 1.7 dB between the capacities $C^2 = C_H^2 = 0.5$ can be observed. At level 1, symbol x^1 is represented double in the underlying 4ASK signal set. Here, the gap between the capacities $C^1 = C_H^1 = 0.5$ reduces to 1.1 dB. This gap is further reduced to 0.9 dB at level 0. Thus, for transmission where symbols are multiply represented by signal points the gap between soft and hard decision decoding becomes smaller. We interpret this different behavior by the following observation: For binary antipodal signaling the gain of soft decision decoding mainly results from those high reliable symbols which are received far away from the (hard) decision boundary. But for multiamplitude/multiphase modulation the existence of boundaries of the decision regions on all sides of inner points rather excludes such points and thus reduces the soft decision advantage. Additionally, as usual the gap between capacities C^i and C_H^i decreases for increasing SNR.

Secondly, the cases SSS and SHH are compared. It is apparently from Fig. 2 that the gap between C and C_{SHH} is negligible for rates above 2.5 bit/symbol. In particular, we focus on the rate 2.5 bit/symbol, since 0.5 bit redundancy per dimension is sufficient to achieve near optimum performance as long as equiprobable signal points are used, cf. [FHW98, WFH97c, FHW96, Ung82]. For $R = 2.5$ bit/symbol the loss for the case SHH vs. the optimum case SSS is 0.13 dB. Even for rates down to 2.0 bit/symbol only a loss of max. 0.6 dB is visible. Hence, the performance loss for a coded modulation scheme due to hard decision decoding at higher levels is dramatically reduced when compared

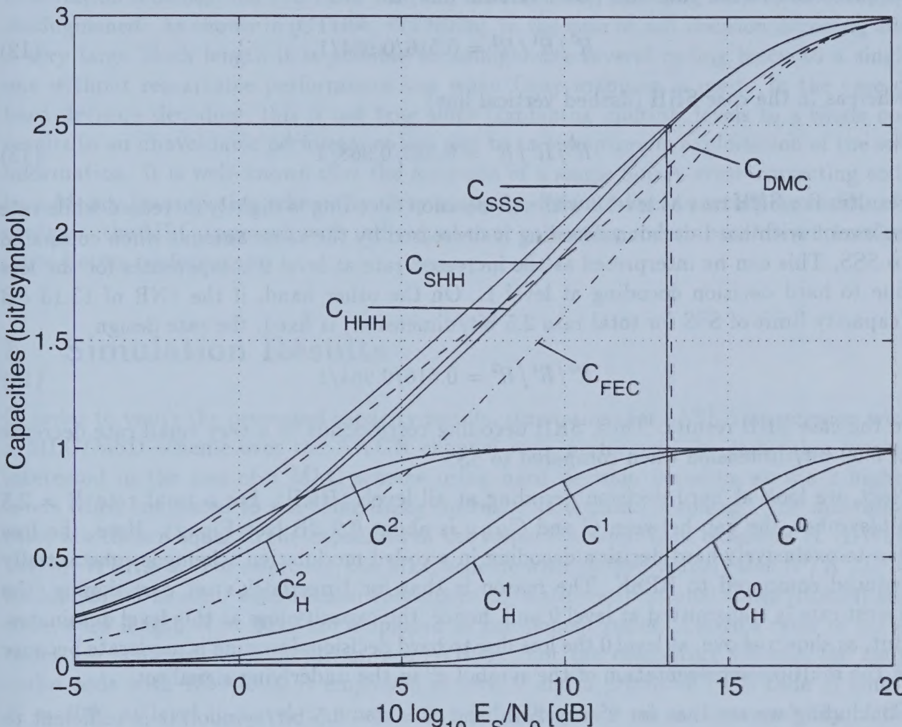


Figure 2: Capacities of 8–ary ASK over the AWGN channel for decoding strategies SSS, SHH, and HHH (see text) and corresponding capacities C^i and C^i_H for the equivalent and equivalent binary symmetric channel i, $i = 0, 1, 2$. Dashed line: Capacity 8–ary DMC. Dashed–dotted line: Capacity 8–ary ASK with FEC (Gray mapping). Solid vertical line: $C_{SSS} = 2.5$ bit/symbol. Dashed vertical line: $C_{SHH} = 2.5$ bit/symbol.

to the case of BPSK. The explanation is as follows. For set partitioning according to Ungerböck's criterion, the individual rates increase from lowest level to highest level. Thus, the performance loss due to hard decision decoding decreases. Hence, if hard decision decoding is only used at higher levels, where high–rate codes are active, the loss remains small.

An additional loss occurs if maximum likelihood decoding is replaced by bounded minimum distance decoding. This loss cannot be assessed exactly, but binary codes with rate very close to 1 are not far away from the sphere–packing bound and hence the difference between bounded minimum distance decoding and hard decision maximum likelihood decoding is not essential. In the case SHH, where only high–rate codes are used for hard decision decoding, there is only a little additional loss for practical bounded minimum distance decoding algorithms.

If we design the system for a total rate $R = 2.5$ bit/symbol, the rate distribution according to capacities in the case SSS (solid vertical line) is

$$R^0/R^1/R^2 = 0.516/0.984/1, \tag{12}$$

whereas in the case SHH (dashed vertical line)

$$R^0/R^1/R^2 = 0.532/0.968/1 \tag{13}$$

results. For SHH rate at level 0 with soft decision decoding is slightly increased while rate at level 1 with hard decision decoding is decreased by the same amount when compared to SSS. This can be interpreted as the increased rate at level 0 compensates for the loss due to hard decision decoding at level 1. On the other hand, if the SNR of 13.13 dB (capacity limit of SSS for total rate 2.5 bit/dimension) is fixed, the rate design

$$R^0/R^1/R^2 = 0.516/0.964/1 \tag{14}$$

in the case SHH results. Thus, SHH decoding corresponds to a very small rate decrease of 0.02 bit/dimension when compared to SSS.

Next, we look at hard decision decoding at all levels (HHH). For a total rate $R = 2.5$ bit/symbol, the gap between C and C_{HHH} is about 0.9 dB (see Fig. 2). Here, the loss due to exclusively hard decision decoding in a coded modulation scheme is substantially reduced compared to BPSK. The reason is that for Ungerböck's set partitioning, the lowest rate is transmitted at level 0 and, hence, the capacity loss at this level dominates. But, as shown above, at level 0 the loss due to hard decision decoding is moderate because of the multiple representation of the symbol x^0 in the underlying signal set.

Concluding we see that for 2^ℓ–ary ASK and $R = \ell - 0.5$ bit/symbol, it is sufficient to employ soft decision decoding only at level 0. Hard decision decoding at higher levels can be done without any remarkable performance loss while offering a reasonable reduction in complexity. Even if hard decision decoding is used at all levels the capacity loss compared to soft decision decoding is lower than 1 dB. Notice that in the interesting region near $R = \ell - 0.5$ bit/dimension the curve C_{FEC} shows exactly the same gap of 1.8 dB to the capacity C of the AWGN channel as in the case of binary antipodal signaling (i.e. $\ell = 1$). The difference between C_{DMC} and C_{FEC} is due to the memory destroying process of interleaving.

Finally, the differences in capacities C_{DMC} (8–ary one–level coding for the 8–ary DMC), C_{FEC} (binary one–level coding for 8ASK, Gray mapping, and bit interleaving) and C_{HHH} (MLC/MSD with hard decision decoding at each level) are discussed. It can be observed from Fig. 2 that the MLC approach with hard decision decoding at each level substantially outperforms the one–level coding schemes. The difference between the schemes is the manner how the soft output of the underlying AWGN channel is exploited. For the 8–ary DMC as well as for one–level FEC the soft channel output is quantized once. In the staged decoding of a MLC scheme, the quantization of the soft value at each level in order to perform the respective binary detection is controlled by the decoding results at lower levels. This leads to the important result that, in the case of hard decision decoding, the most efficient way to exploit the channel information is to split up the coding and decoding procedure in multiple, ideally binary, levels and, thus, to use the soft channel output multiply without the need of complex soft decoding algorithms. That means, 1 dB gain is simply possible by quantizing the channel output step–by–step without any additional complexity. Therefore, dealing with design proposals for coded modulation schemes, the two cases soft and hard decision decoding have to be carefully distinguished. As shown in [CTB96, WFH97b], in the case of soft decision decoding and a very large block length it is possible to amalgamate several coding levels to a single one without remarkable performance loss when Gray mapping is used. In the case of hard decision decoding, this is not true since combining multiple levels to a single one results in an unavoidable performance loss due to an suboptimum exploitation of the soft information. It is well-known that the adoption of a single binary error-correcting code to a M-ary transmission scheme by Gray mapping does not lead to satisfying results. In contrast. the MLC approach with individual error correcting codes at each level promises quite better performance.

5 Simulation Results

In order to verify the presented capacity results, simulations for 8ASK transmission with a MLC/MSD scheme over the AWGN channel were performed. In particular, we are interested in the loss of a MLC scheme using hard decision decoding at the 2 higher levels when compared to a scheme using entirely soft decision decoding. The individual rates are chosen equal to the capacities of the respective equivalent channels, cf. [HW94, WH95, WFH97a]. For reference, the MLC scheme with the individual rates $R^0/R^1/R^2 = 0.52/0.98/1.0$ is used, where turbo codes over 16-state convolutional codes [Rob94] and code word length $N = 4095$ are employed as component codes. Flexible rates of turbo codes are achieved via puncturing, see [WH95]. For the competing scheme, the same turbo code with $R^0 = 0.52$ is employed at level 0 and a primitive BCH code of length $N = 4095$ at level 1. Level 2 remains uncoded. The error correcting capability of the BCH code is adapted in such a way that the individual performances at level 0 and level 1 are similar. As a result, the required error correcting capability for the BCH code is $t = 14$ errors and, hence, $R^1 = 0.96$, which is exactly equal to the capacity C^1 in (14). The simulation results are depicted in Fig. 3. For the reference scheme $10\log_{10}(E_b/N_0) = 10.1$ dB is required to achieve BER $= 10^{-5}$ with a total rate $R = 2.5$ bit/symbol. Since the capacity $C = 2.5$ bit/symbol is reached for $10\log_{10}(E_b/N_0) = 9.1$ dB, this scheme works about 1.0 dB above capacity. From Fig. 3, it can be seen that the competing scheme with the BCH code at level 1 achieves BER $= 10^{-5}$ at $10\log_{10}(E_b/N_0) = 10.15$ dB with a total rate $R = 2.48$ bit/symbol. Here, the capacity $C = 2.48$ bit/symbol is reached for $10\log_{10}(E_b/N_0) = 8.9$ dB. Hence, the MLC scheme using a BCH code at level 1 works about 1.25 dB above capacity, resulting in a loss of about 0.25 dB vs. the MLC scheme using turbo codes at both coded levels. This loss of 0.25 dB obtained by simulations fairly corresponds to the loss 0.15 dB predicted from capacity argument.

6 Conclusions

Employment of hard decision FEC decoding at some coding levels was proposed as a very efficient method to save complexity in coded modulation schemes. For Ungerböck's labeling of signal points, the performance loss compared to soft decision decoding (in terms of capacity as well as in simulations) is only about 0.13 dB for 8ASK transmission over the AWGN channel, when hard decision decoding is employed at all but the lowest

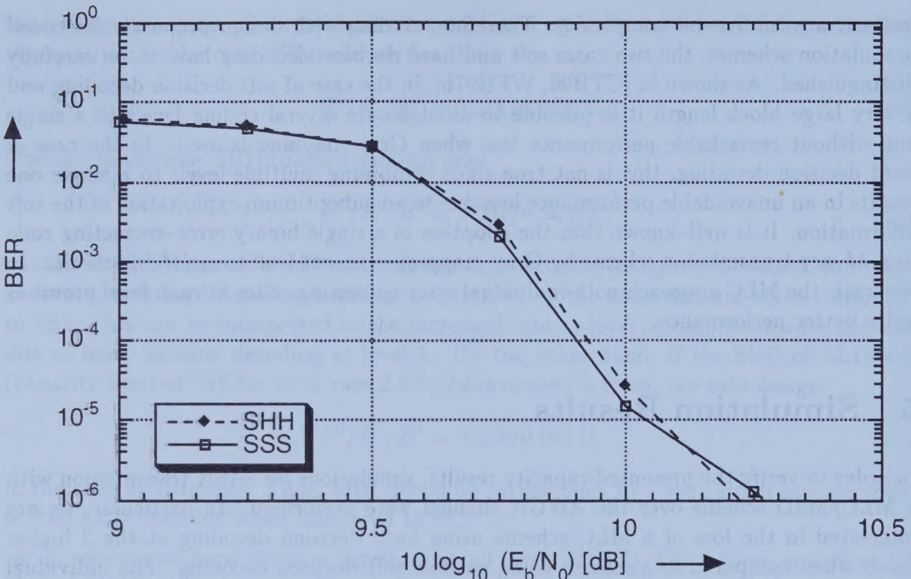


Figure 3: BER of 8ASK transmission with MLC/MSD over the AWGN channel. Code word length $N = 4095$. Solid line (case SSS): Rate distribution $R^0/R^1/R^2 = 0.52/0.98/1.0$, total rate $R = 2.5$ bit/symbol. Turbo codes are employed as component codes. Dashed line (case SHH): Rate distribution $R^0/R^1/R^2 = 0.52/0.96/1.0$, total rate $R = 2.48$ bit/symbol. Component codes: Turbo code at level 0, 14–errors–correcting BCH code at level 1. Simulation results.

level. Since in general the effort for hard decision decoding is substantially lower than for soft decision decoding, we can state that power efficient coded 2^ℓ–ary, $\ell > 1$, modulation requires only slightly additional complexity when compared to coded binary transmission. Thus, bandwidth efficient digital transmission with a multiamplitude/multiphase constellation close to capacity limit requires much less decoding complexity per bit than binary antipodal signaling. If hard decision decoding is employed at all levels, a MLC approach with ℓ individual binary error–correcting codes is recommended. Since staged decoding exploits the soft channel output to some extent, it promises quite better performance than a single binary code adapted to the 2^ℓ–ary modulation scheme with Gray mapping.

Acknowledgment

The authors are greatly indebted to *Deutsche Forschungsgemeinschaft* for supporting this work. Additionally, the authors like to thank Rolf Weber [Web97] for performing some of the calculations in this paper.

References

[CTB96] G. Caire, G. Taricco, and E. Biglieri. Capacity of bit–interleaved channels. *Electronics Letters*, vol.32:pp.1060–1061, June 1996.

[FHW96] R. Fischer, J. Huber, and U. Wachsmann. Multilevel coding: Aspects from information theory. In *Proc. CTMC at IEEE Global Telecommun. Conf. (GLOBECOM'96)*, pages 26–30, London, UK, Nov. 1996.

[FHW98] R. Fischer, J. Huber, and U. Wachsmann. On the Combination of Multilevel Coding and Signal Shaping. Submitted to 2nd ITG Conf. Source and Channel Coding, Aachen, Germany, March 1998.

[HW94] J. Huber and U. Wachsmann. Capacities of Equivalent Channels in Multilevel Coding Schemes. *Electronics Letters*, vol. 30:pp. 557–558, March 1994.

[IH77] H. Imai and S. Hirakawa. A new multilevel coding method using error correcting codes. *IEEE Trans. Inf. Theory*, vol.23:pp.371–377, May 1977.

[KZS94] Y. Kofman, E. Zehavi, and S. Shamai (Shitz). Performance Analysis of a Multilevel Coded Modulation System. *IEEE Trans. Commun.*, vol.42:pp.299–312, 1994.

[PT89] G.J. Pottie and D.P. Taylor. Multilevel Codes based on Partitioning. *IEEE Trans. Inf. Theory*, vol.35:pp.87–98, Jan. 1989.

[Rob94] P. Robertson. Illuminating the Structure of Code and Decoder for Parallel Concatenated Recursive Systematic (Turbo) Codes. In *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, pages 1298–1304, San Francisco, Nov./Dec. 1994.

[Ung82] G. Ungerböck. Channel coding with multilevel/phase signals. *IEEE Trans. Inf. Theory*, vol.28:pp.55–67, Jan. 1982.

[Web97] R. Weber. *Einsatz von fehlerkorrigierenden Block–Codes in der codierten Modulation.* Diploma thesis, Lehrstuhl für Nachrichtentechnik, Universität Erlangen-Nürnberg, Germany, 1997. In German.

[WFH97a] U. Wachsmann, R. Fischer, and J. Huber. Multilevel Coding: Basic Concepts, Capacity, and Rate Design. Submitted to IEEE Trans. Inf. Theory, 1997.

[WFH97b] U. Wachsmann, R. Fischer, and J. Huber. Multilevel Coding: Dimensionality of the Constituent Signal Set, Labeling, and Hard Decision Decoding. Submitted to IEEE Trans. Inf. Theory, 1997.

[WFH97c] U. Wachsmann, R. Fischer, and J. Huber. Multilevel Coding: Distance Profile and Signal Shaping. Submitted to IEEE Trans. Inf. Theory, 1997.

[WH95] U. Wachsmann and J. Huber. Power and bandwidth efficient digital communication using turbo codes in multilevel codes. *Europ. Trans. Telecommun. (ETT)*, vol. 6:pp. 557–567, Sept.-Oct. 1995.

Ellipsoids in Hamming spaces: isoperimetry, random search, and maximum likelihood decoding*

Ilya Dumer

College of Engineering, University of California, Riverside, CA 92521

1 Introduction

In this paper, we study soft-decision maximum-likelihood decoding algorithms that try to find an error-free information subset in a number of trials. Given a soft-decision output y in the Euclidean space R^n, the transmitted codeword is necessarily retrieved whenever it is the most probable one. The main issue is to design an algorithm that performs this task in a minimum possible number of trials. The special case is complete minimum distance (MD) decoding on hard-decision binary symmetric channels (BSCh). This problem has been extensively studied, and the results are as follows.

Let $C(n, R)$ be a binary linear code of length n and code rate R in the Hamming space E_2^n and let D be the corresponding Gilbert distance

$$D = \max\{d : \sum_{i=0}^{d-1} \tbinom{n}{i} \leq 2^{n(1-R)}\}. \tag{1}$$

Also, let $S(n, t)$ be the set of vectors of Hamming weight t in E_2^n. For a given output $y \in E_2^n$, MD decoding should retrieve the closest codeword in the Hamming metric no matter how far it is. However, Evseev has proved [12] that correction of only D or fewer errors at most doubles the decoding error probability of MD decoding for *any* linear (n, R) code. Another important result by Blinovskii [2] shows that *virtually all long* linear (n, R) codes have *covering radius* $\rho = D + O(\ln n)$. In other words, correction of slightly more than D errors necessarily provides full MD decoding for virtually all long linear codes. So, MD decoding needs to correct only about 2^{n-k} most probable error patterns located within the sphere $S(n, D)$.

---

*This research was supported by the NSF grant NCR-9703844.

These results were efficiently used [13], [5] in designing MD decoding algorithms that seek error-free sets of $k = nR$ positions. Indeed, we need to find only k error-free information positions among at least $n - D$ error-free positions. In other words, we can try to find any set of $n - k$ positions that includes all D erroneous positions. More formally, we need to find a *covering set* $T(n, n - k, D) \subset S(n, n - k)$ such that any combination of D positions is necessarily covered by at least one vector $v \in T$. The lower and upper bounds on the minimal size $|T_{\min}|$ are well known [11]:

$$\binom{n}{D} \Big/ \binom{n-k}{D} \leq |T_{\min}(n, n-k, D)| \leq \left[\ln \binom{n-k}{D} + 1\right] \binom{n}{D} \Big/ \binom{n-k}{D}. \tag{2}$$

The first important point is that these two bounds coincide up to the exponential order $L_0 = \binom{n}{D} / \binom{n-k}{D}$. Second, the upper bound is achieved by a simple random search. Namely, we perform some number L of trials choosing equally likely any vector $v \in S(n, n-k)$ in any trial. Then it is easily verified [11] that all the vectors of weight D are covered with probability $P(L)$ that is strictly separated from 0 for any $L \geq [\ln \binom{n-k}{D} + 1]L_0$. Moreover, once about $n \ln nL_0$ trials are performed, we fail to get a covering with a negligible probability $\exp\{-n \ln n\}$ that declines faster than the probability of incorrect decoding. So one can either perform this search in advance or incorporate it in decoding directly. In every trial, we get $n-k$ positions and re-encode the symbols in the remaining (presumably, error-free) k positions. Each codeword is compared with the received vector v and the closest codeword is chosen. The only question to be solved is whether we can re-encode any k-subset as an information one. This in general cannot be done. However, it is easily proved that for virtually all linear codes all k *– subsets* are *quasi-information* ones. The latter means that only about $\exp\{\sqrt{n}\}$ or fewer codewords can have the same set of k symbols on any k-subset. So, the above number $\binom{n}{D} / \binom{n-k}{D}$ defines the complexity 2^{nc} of MD decoding. The exponent c is easily verified to be $c = (1 - R)(1 - H(\frac{\delta}{1-R}))$, where $\delta = H^{-1}(1 - R)$ is the relative GV distance D/n. Note that for the rate $R = 1/2$ this exponent c is four times less than the trellis complexity $c_0 = \min(R, 1 - R)$. It is slightly reduced further in [9] by combining the above technique with a presorting procedure from [8].

For soft decision maximum-likelihood (ML) decoding the problem is much different. Any received output $y \in R^n$ gives the most probable "hard decision" vector $a \in E_2^n$. However, each symbol a_j has its own reliability v_j, and it is clear that different subsets of $n - k$ positions are not equally reliable. Nor do the sets of most probable error patterns form a sphere. In general, we need to find an algorithm that not only chooses ($n - k$) covering subsets in the least number of trials but does so for any set of n reliabilities v_j. More specifically, the following questions arise:

- How many error patterns need to be tested in order to get explicit ML decoding ?

- Which error patterns should be included in the minimal list ?
- How can these error patterns be covered ?
- How many trials do we need to get the full covering ?

The answer to the first problem was recently given in [10]. It is proved that we need to test the same order of 2^{n-k} most probable error patterns as we do in MD decoding. As we will see below, these error patterns form an ellipsoid of size about 2^{n-k} with arbitrary set of axial coefficients $v = (v_1, \ldots, v_n)$ defined by received reliabilities. So, our main problem is to design an algorithm that produces a minimum covering for any ellipsoid of size 2^{n-k}, the above set v being given. This problem is solved in this paper. It turns out that for any ellipsoid, its minimal covering includes at most the same order $\binom{n}{D} / \binom{n-k}{D}$ of $(n-k)$ subsets as that needed in its spherical $T(n, n-k, D)$ counterpart. More specifically, we design a random search algorithm that upper-bounds the minimum covering size by the above exponential order $c = (1-R)(1 - H(\frac{\delta}{1-R}))$ for any ellipsoid. In doing so, we choose each position j with some probability β_j independently of other positions. To get $n-k$ or fewer positions chosen, we use the equality $\beta_1 + \ldots + \beta_n = n-k$. Then we eliminate all the trials, where the number of chosen positions exceeds $n-k$. The values $\beta_1, \ldots \beta_n$ are optimized depending on symbol reliabilities. Interestingly, most ellipsoids need the coverings that are *exponentially smaller, so that the spherical covering gives the worst case.* We also consider the lower bounds on the minimum covering and prove that for a variety of ellipsoids the upper bounds are tight. The end result is the explicit ML decoding algorithm that gives the above complexity exponent c for virtually all linear codes even in the worst case. Our conjecture is that the obtained upper bounds on minimum coverings are tight for all ellipsoids, and an open problem is to obtain the tight lower bounds.

In the next section we examine the sets of the most probable inputs as ellipsoids in E_2^n, and study isoperimetric problems for arbitrary ellipsoids. Then in Section 3 we present a random-search algorithm that produces a minimal exponential covering for any ellipsoid. Section 4 concludes our discussion.

2 Ellipsoids in Hamming spaces and isoperimetric problems

Ellipsoidal coverings are important due to their close relation to maximum likelihood (ML) soft decision decoding. First, let a binary code $C(n, R)$ be used over a memoryless channel $(E_2^n, Y^n, P(y|x))$. The latter is defined by the input space E_2^n, an output space Y^n that can be infinite, and conditional probability $P(y|x) \overset{def}{=} \prod_{j=1}^{n} p(y_i|x_i)$ of receiving an output $y \in Y^n$ for a given $x \in E_2^n$. In the sequel, we suppose that all 2^{nR} messages are equiprobable. Then for a given output y, ML decoding $\psi(y) = c$ retrieves a codeword $c \in C$ with the maximum *a posteriori* probability $P(c|y) \geq P(c'|y), c' \neq c$.

Second, let $a = (a_1, .., a_n)$ be the most probable input for a given output y. The latter means that $p(a_j|y_j) \geq 1/2$ for all $j = 1, ..., n$. For any position j, we relate the likelihood of the more probable symbol a_j to the less probable (complementary) symbol $x_j = 1 - a_j$ by the number $v_j = \ln p(a_j|y_j) - \ln p(x_j|y_j)$. Then the a posteriori probability $P(x|y)$ of any input $x \in E_2^n$ is

$$\ln P(x|y) = \ln P(a|y) - \sum_{j:x_j \neq a_j} v_j = \ln P(a|y) - \sum_{j=1}^{n} v_j(x_j - a_j)^2.$$

Now for a given vector $v = (v_1, ..., v_n)$ with n positive real components, we define the "ellipsoidal distance" $w_v(x, a) = \sum_j v_j(x_j - a_j)^2$ between the points $x, a \in E_2^n$. Then we consider *the ellipsoid*

$$E(a, v, r) = \{x \in E_2^n | \sum_j v_j(x_j - a_j)^2 \leq r\} \tag{3}$$

centered at any point $a \in E_2^n$. Obviously, this forms the set of $M(v, r) = |E(a, v, r)|$ most probable input vectors. In terms of ellipsoidal distance, $E(a, v, r)$ is the set of $M(v, r)$ lightest inputs $x \in E_2^n$. For a given v, let $E(a, v, M)$ denote the smallest ellipsoid that includes at least M points, and r_M be the corresponding radius. Below, we use both notations $E(a, v, M)$ and $E(a, v, r_M)$. Now consider the ellipsoids of *exponential* size 2^{n-k} using the following definition.

Definition 1 *We say that the functions $M(n)$ and $N(n)$ have the same exponential order and write $M(n) \overset{\exp}{=} N(n)$, if $\lim_{n\to\infty} \log[M(n)/N(n)] = 0$.*

Definition 2 *We say that a code $C(n, R)$ is M-good, if every ellipsoid $E(a, v, M)$ includes codeword(s) from C regardless of $a \in E_2^n$ and v.*

The following theorem justifies why we study the ellipsoids of exponential size 2^{n-k}.

Theorem 1 *[10]There exists $M_0 \overset{\exp}{=} 2^{n(1-R)}$ such that virtually all linear (n, R)-codes are M-good so that any ellipsoid $E(a, v, M)$ includes codewords.*

Corollary 1 *[10]There exists $M_0 \overset{\exp}{=} 2^{n(1-R)}$ such that ML decoding can be performed only within the ellipsoid $E(a, v, M_0)$ for virtually all long linear codes $C(n, R)$.*

In other words, the most probable codeword necessarily belongs to the list of about 2^{n-k} most probable inputs, which in essence form the ellipsoid $E(a, v, 2^{n-k})$ centered at the hard-decision vector $a \in E_2^n$. Recently, these discrete ellipsoids have been extensively studied. In particular, Zuev[19] has proved that the number of different ellipsoids in E_2^n is lower- and upper-bounded by the orders $2^{n^2(1-10/\ln n)}$ and 2^{n^2}. Pinsker [16] has found the explicit asymptotic exponent $[\log_2 M(v, r)]/n$ as $n \to \infty$. We first study the relationship between the size M and ellipsoidal radius r. Without loss of generality we re-scale the reliabilities v_i so that $\sum v_i = n$. This allows us to relate the parameters of all ellipsoids to the sphere $S(n, D)$, for which $v_1 = \ldots = v_n = 1$.

Lemma 1 *For any ellipsoid* $E(a. v. 2^{n-k})$ *the asymptotic inequality* $r \preceq D + O(\ln n)$ *holds as* $n \to \infty$.

This lemma shows that the sphere $S(n, D)$ has the largest radius among all ellipsoids of the same size 2^{n-k}. Now let $s = \max\{w_H(x) | x \in E(a, v. 2^{n-k})\}$ be the largest Hamming weight among all the points of the ellipsoid. We call this the *Hamming radius* of ellipsoid. Obviously, $s \succeq D$, since $|S(n, s)| \overset{\exp}{<} 2^{n-k}$ if $s/D < 1$.

Lemma 2 *For any ellipsoid* $E(a, v, 2^{n-k})$ *of Hamming radius* s *the ellipsoidal radius* r *is upper-bounded as*

$$r \preceq n\frac{\delta_s}{1/2+\delta_s} + O(\ln n), \text{ where } \delta_s = H^{-1}(1 - \frac{R}{1-s/n})$$

Note that the relative GV distance is $\delta = \delta_0$ and that $\delta_0 > \delta_1 > \ldots > \delta_{n-k} = 0$. In other words, this implies that the greater the Hamming radius s of the ellipsoid the smaller its ellipsoidal radius r is. The sphere $S(n, D)$ has the smallest Hamming radius D and the biggest ellipsoidal radius D among all ellipsoids of exponential size 2^{n-k}.

Example 1 *Cube as an ellipsoid.*

Consider the reliability set $v_1 = \ldots = v_{n-k} = \frac{n}{(n-k)(k+2)}$, and $v_{n-k+1} = \ldots = v_n = \frac{n(k+1)}{(n-k)(k+2)}$. Then $\sum_1^n v_j = 1$ whereas $\sum_1^{n-k} v_j < v_{n-k+1}$. The latter implies that the ellipsoid $E(0, v, 2^{n-k})$ forms the cube $B_{n-k} = \{(x_1, \ldots, x_n) \subset E_2^n | \, x_{n-k+1} = \ldots = x_n = 0\}$ generated by the first $n-k$ coordinates. Its ellipsoidal radius is $r = n/(k+2)$ and the Hamming radius is $s = n-k$. This ellipsoid formed by two groups of sufficiently different reliabilities presents one of the extreme cases, the Hamming sphere being the other one. ☐

Our next question is how to find the ellipsoidal weight spectrum within an arbitrary ellipsoid $E(a, v, r)$. For given $\varepsilon > 0$, we consider the subset $E_\varepsilon(a, v, \rho) = \{x | \, \rho(1-\varepsilon) \leq w(x) \leq \rho(1+\varepsilon)\}$ of points whose ellipsoidal weights cover the interval $[\rho(1-\varepsilon), \rho(1+\varepsilon)]$.

Below this is called the *(v, ρ, ε)-layer*. Our problem is to find the size $|E_\varepsilon(a, v, \rho)|$. This part of the work closely follows the work by M. Pinsker[16] who calculated the explicit exponential size for any ellipsoid. We arrive at the similar results from a different prospective. For a given $\rho < n$, define the parameters $\lambda_\rho > 0$ and $p_j \leq 1/2$ by the following equations

$$\sum_j \frac{v_j}{1 + 2^{\lambda_\rho v_j}} = \rho, \quad p_j = p_j(\rho) = \frac{v_j}{1 + 2^{\lambda_\rho v_j}}. \tag{4}$$

These parameters are of major importance in designing good coverings for arbitrary ellipsoids.

Lemma 3 *For $n \to \infty$ and $\rho \to \infty$, the size of any (v, ρ, ε)-layer $E_\varepsilon(a, v, \rho)$ satisfies the asymptotic inequality*

$$2^{n \sum H(p_j) - \rho \varepsilon \lambda_\rho} \preceq |E_\varepsilon(a, v, \rho)| \preceq 2^{n \sum H(p_j) + \rho \varepsilon \lambda_\rho}.$$

In other words, any (v, ρ, ε)-layer has exponential size $2^{n \sum H(p_j)}$. Now consider the ellipsoid $E(a, v, r)$, and let μ and $\lambda = \lambda_\mu$ be the parameters such that

$$\sum H(p_j(\mu)) = \max\{\sum H(p_j(\rho)) | \rho \leq r\}. \tag{5}$$

In other words, the *(v, μ, ε)-layer is the "heaviest"* one and gives the exponential size $2^{n \sum H(p_j(\mu))}$ of the whole ellipsoid. It can be proved that for all but exceptional cases $\mu = r$. However, Example 1 shows that it is not necessarily so. In this case, the boundary layer includes only one point with ones on the first $n - k$ coordinates whereas the most numerous layer includes all the vectors of Hamming weight $(n-k)/2$ on these coordinates.

Finally, we study the Hamming weight distribution within the ellipsoid $E(a, v, r)$. Similarly, we define the *spherical (v, s, ε)-layer* as the subset $E_\varepsilon(a, v, s) = \{x| \ s(1 - \varepsilon) \leq w_H(x) \leq s(1 + \varepsilon)\}$ of points whose Hamming weights are about s. For given v, μ and λ, we define

$$t = \sum p_j(\mu)) = \sum_j \frac{1}{1 + 2^{\lambda v_j}}. \tag{6}$$

Lemma 4 *The spherical (v, t, ε)-layer has the same exponential size $2^{n \sum H(p_j(\mu))}$ as the whole ellipsoid.*

Lemma 5 *For any ellipsoid the inequality $D \leq t \leq (n - k)/2$ holds.*

Note that the lower bound $t = D$ is achieved on the sphere $S(n, D)$, and the upper one is met for the cube from Example 1. More detailed study shows that most ellipsoids $E(a, v, r)$ include numerous points whose Hamming weights considerably exceed the above weight t. Therefore we need to cover the most numerous layer of Hamming weight t as well as many other points of higher weights. For example, only one $n-k$ subset generated by the first $n-k$ coordinates covers the whole cube from Example 1. At the first sight, this seems to give the covering size at least $\binom{n}{t} / \binom{n-k}{t}$. This exceeds the spherical covering size $\binom{n}{D} / \binom{n-k}{D}$ for $t > D$. However, more precise study shows that the situation is very different. It turns out that the covering size can be about $\binom{n}{D} / \binom{n-k}{t}$ that is less than $\binom{n}{D} / \binom{n-k}{D}$ for $t > D$. Loosely speaking, the number 2^{n-k} of points to be covered matters much more than their weights. However, the precise analysis is much more involved, and we only formulate the main results below.

3 Main results

We start with the upper bounds on ellipsoidal coverings. The main idea is to use different probabilities $\beta_1, ..., \beta_n$ for different positions $j = 1, ..., n$ and choose these positions independently every time while seeking vectors of Hamming weight $n-k$. More precisely, we choose "1" in position j with probability β_j and "0" with the complementary probability $1-\beta_j$. The first constraint is obvious:

$$\beta_1 + ... + \beta_n = n - k.$$

In this case, most designed vectors have Hamming weight about $n-k$, and sufficiently many can be used in our covering. Second, we use the constraint

$$\beta_j \geq 2p_j(\mu),$$

where $p_j = p_j(\mu)$ are defined in (5). This restriction can be satisfied since $p_j < 1/2$. Now the main result follows.

Theorem 2 *For given $\beta_1, ..., \beta_n$, the ellipsoid $E(0, v, 2^{n-k})$ can be covered by at most*

$$L(\beta_1, ..., \beta_n) = (n \ln n) 2^{\sum [n-k-\beta_j H(p_j/\beta_j)]}$$

randomly chosen vectors of Hamming weight $n-k$ with probability $1 - \exp\{-n \ln n\}$.

Theorem 3 *Let the probabilities p_j satisfy the inequality*

$$p_j \leq t/(n-k) \tag{7}$$

for all $j = 1, ..., n$, where t is defined in (6). Then

$$\min L(\beta_1, ..., \beta_n) \overset{\exp}{=} 2^{n-k} / \tbinom{n-k}{t}. \tag{8}$$

Theorem 4 *For an ellipsoid $E(0, v, 2^{n-k})$ the minimal covering size is bounded from below by $L = 2^{n-k}/\binom{n-k}{t}$ vectors of weight $n-k$.*

In other words, the lower and upper bounds on ellipsoidal covering coincide if the restrictions (7) hold. Finally, the following theorem shows that the size of ellipsoidal covering is upper-bounded by its spherical counterpart.

Theorem 5 *Any ellipsoid $E(0, v, 2^{n-k})$ can be covered by at most*

$$L = (n \ln n) 2^{\sum H(p_j)[1-H(\frac{p_j}{H(p_j)})]} \overset{\exp}{\leq} \binom{n}{D} / \binom{n-k}{D} \tag{9}$$

randomly chosen vectors of Hamming weight $n-k$ with probability $1 - \exp\{-n \ln n\}$.

Corollary 1 *For virtually all linear binary (n, R)-codes used on any memoryless channel, ML decoding complexity is upper-bounded by exponential order 2^{nc}, where $c = (1-R)(1 - H(\frac{\delta}{1-R}))$.*

4 Concluding remarks

A few concluding remarks seem to be in order. The above bounds (8) and (9) define the maximum complexity in the worst case. The average complexity is smaller for any code rate R below the channel capacity. It is also interesting that the average complexity will slightly decline as the rate R approaches the capacity. Finally, for the lengths $n = 64$ and 128, we compare complexity of random-search maximum-likelihood decoding with the general upper bound $2^{n \min(R, 1-R)}$ on trellis complexity [1], [18] and with its refinement for BCH codes of length $n = 64$ [17]. Random-search complexity is upper bounded by its hard-decision counterpart (2). These are summarized in the following two tables. Various complexities 2^{nc} are presented by the logarithmic scale nc and are rounded off.

R	nmin(R,1-R)	BCH Trellis	Covering
0.28	18	17	11.4
0.38	24	16	11.9
0.47	30	21	11.7
0.61	25	20	10.5
0.70	19	14	9.0
0.80	13	12	6.7

Table 1. Complexity for $n = 64$

R	nmin(R,1-R)	Covering
0.28	36	19.5
0.38	49	20.5
0.47	60	20.5
0.61	49	18.7
0.70	38	16.3
0.80	26	12.7

Table 2. Complexity for $n = 128$

The author wishes to thank R. Krichevskii, C. Johnson and R. Redmon for discussions and help in computer simulation.

References

[1] L.R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory,* vol. 20, pp. 284-287, 1974.

[2] V.M. Blinovskii, "Lower asymptotic bound on the number of linear code words in a sphere of given radius in F_q^n," *Probl. Peredachi Inform.*, vol. 23, no. 2, pp. 50-53, 1987 (in Russian).

[3] V.M. Blinovskii, "Asymptotically exact uniform bounds for spectra of cosets of linear codes," *Probl. Peredachi Inform.*, vol. 26, no. 1, pp. 99-103, 1990.

[4] D. Chase, "A class of algorithms for decoding block codes with channel measurement information," *IEEE Trans. Inform. Theory,* vol. 18, pp. 170-182, 1972

[5] J.T. Coffey and R.M.F. Goodman, "The complexity of information set decoding", *IEEE Trans. Inform. Theory,* vol. 36, pp. 1031-1037, 1990.

[6] J.T. Coffey, R.M.F. Goodman, and P.G. Farrell, "New approaches to reduced-complexity decoding", *Discrete Appl. Mathematics,* vol. 33, pp. 43-60, 1991.

[7] I. Dumer, "Suboptimal decoding of linear codes. Partition technique," *IEEE Trans. Inform. Theory,* vol. 42, Nov. 1996, pp. 1971-1986.

[8] I.I. Dumer, "Two algorithms for linear codes decoding", *Probl. Peredachi Inform.*, vol. 25, no. 1, pp. 24-32, 1989 (in Russian).

[9] I.I. Dumer, "On minimum distance decoding of linear codes", *In: Proc. Fifth Intern. Workshop on Information Theory "Convolutional Codes; Multiuser Communication"*, Moscow, pp. 50-52, Jan. 1991.

[10] I. Dumer, "Covering lists in maximum likelihood decoding," *Proc. 34 Annual Allerton Conf. on Commun., Control, and Comp.*, pp. 683-692, Monticello, IL, Oct. 2-4, 1995.

[11] P.Erdos and J.Spencer, *Probabilistic methods in combinatorics.* Akademiai Kiado, Budapest, 1974.

[12] G.S. Evseev, "On the complexity of decoding linear codes", *Probl. Peredachi Inform.*, vol. 19, no. 1, pp. 3-8, 1983 (in Russian).

[13] E.A. Krouk, "A bound on the decoding complexity of linear block codes", *Probl. Peredachi Inform.*, vol. 25, no. 3, pp. 103-106, 1989 (in Russian).

[14] A. Lafourcade and A. Vardy, "Lower bounds on trellis complexity of block codes," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1938-1952, 1995.

[15] L. Levitin and C. Hartmann, "A new approach to the general minimum distance decoding problem: The zero-neighbors algorithm", *IEEE Trans. Inform. Theory*, vol. 31, pp. 378-384, 1985.

[16] M.S. Pinsker, Private communication, 1995.

[17] A. Vardy and Y. Be'ery, "Maximum likelihood soft decision decoding of BCH codes," *IEEE Trans. Inform. Theory*, vol. 40, pp. 546-554, 1994.

[18] J.K. Wolf, "Efficient maximum likelihood decoding of linear codes using a trellis", *IEEE Trans. Inform. Theory*, vol. 24, pp. 76-80, 1978.

[19] Yu. Zuev, "Asymptotic logarithm of the number of Boolean threshold functions," *Dokl. Acad. Nayk USSR*, vol. 306, no. 3, pp. 528-530, 1989.

Low-Power Pulse-Shaping Digital Filters

Alan N. Willson Jr.
Electrical Engineering Department
University of California, Los Angeles
willson@ee.ucla.edu

Abstract

We describe techniques for reducing by approximately half (by 2/3 in the linear-phase case) the number of add/subtract operations required by multiplierless pulse-shaping filters for BPSK, QPSK, and QAM modulators.

Introduction

A pulse-shaping digital filter is a system consisting of a data rate expander (that inserts 0's between successive input bits) followed by a digital filter operating at the higher data rate. The system accepts an input stream of symbols which, in the most elementary form, is a stream of one-bit data, and produces an output stream of *B*-bit data (where *B* is a fixed positive integer) having a waveform suitable for processing by a digital-to-analog converter (D/A). The D/A yields an analog signal whose waveform contains the original bit-stream's information while having its frequency content appropriately bandlimited for transmission through some given communication channel. Pulse-shaping filters, both analog and digital, have been well studied in the literature for several decades and their importance in the implementation of practical communication systems has only increased as the interest in transmitting *digital* data has increased. The recent interest in wireless digital communication systems has brought into the picture the notion that the building blocks of portable transceivers, pulse-shaping filters included, must be implemented in a manner that dissipates as little power as possible.

Fig. 1 shows the impulse response of a typical "Nyquist filter." The combined digital transmit and receive filters in a typical communication system are often designed to approximate such filters [1-4]. The Fig. 1 signal is the sort of signal a receiver's A/D converter and digital filters would produce when the transmitted bit-stream consisted of just a single bit, i.e., a 1. If a 0 was transmitted then a sequence having the opposite polarity would be produced within the receiver. A general input bit-stream would produce superimposed copies of the Fig. 1 sequence or its negative, with each copy corresponding to an input bit and, thus, each copy offset (delayed) from the preceding copy by the number of output samples produced per input bit (four, for the Fig. 1 example). Since the Fig. 1 impulse response has equally spaced 0 values on each side of the center point, it's evident that, in this superpositioning process, no other response-sequence copies will interfere with the center value of any input bit's response sequence. This absence of inter-symbol interference provides the means by which, once appropriate synchronization is achieved, the original bit-stream can be recovered.

A transmitter's pulse-shaping filter must have an impulse response that, when cascaded with other filters in the transmitter and receiver, yields a close approximation to the Fig. 1 type impulse response. Its design is also influenced by the requirement that its response sequence, when processed by the transmitter's other digital filters and its D/A, must produce a suitably bandlimited waveform. Techniques for FIR pulse-shaping filter design have been reported [1-

---

This research was supported by NSF Grant MIP-9632698, DARPA Contract F04701-97-C-0010, ONR Grant N00014-95-1-0231, and California MICRO Grant 97-207.

4] and we'll assume that we've designed such a filter, shown in Fig. 2, having N taps, and that we've determined its B-bit tap-weight values. Since this filter operates upon an expanded bit-stream, with the filter considering a -1 value to correspond to each true input-data bit of 0 (as distinct from the 0 bits added by the expander, which must be treated as 0) the filter simply needs to compute a sum of terms consisting of each tap weight multiplied by a member of the set $\{0, 1, -1\}$. That is, just $N - 1$ additions are required per input bit, for an N-tap filter.

While the processing of the expanded input bit-stream could be accomplished exactly as described in the preceding paragraph, there are well-known ways to improve the system's organization and efficiency by combining the expansion and filtering operations into a set of smaller FIR filters that operate in parallel at the lower (unexpanded) data rate of the input bit-stream and with only the 1 and -1 bits occurring. Such "polyphase implementations" are standard and are discussed, for example, in [5]. We thus turn our attention to a typical polyphase component filter, which we can still envision as being represented by the Fig. 2 structure; but now we have an input data stream containing no expander 0's, i.e., consisting of a sequence of points taken from the set $\{1, -1\}$. Since, however, it's highly advantageous to employ one-bit data, and since the actual data stream is comprised of 0 and 1 bits, *that* is what is provided to each polyphase component filter. It will be the filter's responsibility to treat each 0 in its input data stream as though it had the value -1. Thus, the filter produces an output sample by accumulating an algebraic sum of the tap weights, where an h_k value is either added or subtracted depending on whether its data bit corresponds to an input bit of 1 or 0, respectively. Letting N now denote the length of this polyphase component filter, we require $N - 1$ add/subtract operations to produce each output sample. One way to implement the filter, given sufficiently low data rates, would be to sequentially process the tap weights $h_0, \ldots, h_{N-1}$, accumulating their algebraic sum using a single add/subtract unit. However it's organized, some such accumulation process must be performed each time a new input bit is accepted by the Fig. 2 filter. It will next be our objective to arrange for a reduction in this number of add/subtract operations by approximately half.

Reducing add/subtract operations by about half

Clearly, if there were an equal mix of 0/1 bits in the incoming bit-stream and if we could process them directly as 0 and 1, rather than treating the 0 as if it were -1, we could achieve our goal of reducing the number of add/subtract operations by just *skipping* the subtraction operation for all tap weights having a corresponding 0 bit. However, this would obviously give the wrong result for the computed output sample value; moreover, it's not evident that enough 0 input bits would necessarily be present at some particular point in time that any savings would be achieved. Fortunately, we've discovered ways to remedy both of these concerns. In addition to the potential for power savings by requiring fewer add/subtract operations, the processing of one-bit data in the normal manner also provides the design advantages of a simplified system.

Suppose we start with the *bias* or *offset* value $-\Sigma h_k = -(h_0 + \ldots + h_{N-1})$ in the accumulator where the output sample value is computed. Suppose, also, we change each tap weight from h_k to $2h_k$, resulting in the Fig. 3 filter structure. Now, if we process each 1 and 0 input bit normally we'll compute a resulting output value that's equivalent to that produced by the above-described filter wherein each 0 is processed as -1. This fact is obvious because each tap weight that corresponds to an input 1 bit will cause twice its correct weight, i.e., $2h_k$, to be added to the output accumulator; however there will also have been a corresponding $-h_k$ amount present in the sum from the $-\Sigma h_k$ initial bias. Similarly, when a tap weight corresponds to a 0 input bit, nothing's added to the output accumulator as that's exactly what's needed—there's already a $-h_k$ value as a component of the initial bias. Clearly, just by skipping the add/subtract operation for all tap weights h_k corresponding to a 0 input bit, we can save computational effort; that is, power savings can result. If we could only be assured that, in the computation of *each* output bit, at least half the input data bits were 0, then we'd always be assured of saving at least 50% of the computation required.

For many types of communication systems, there are certain processes that the incoming data go through which ensure that the 1 and 0 bits appearing in the pulse-shaping filter's input sequence will occur with equal likelihood. If that situation applies then we can simply build a system employing the above-described technique and we'll, on average, use about half the add/subtract operations, thereby lowering the power dissipation proportionally. Suppose, however, we wanted to actually ensure that at most *N*/2 add/subtract operations were required in the computation of *each* output sample—no matter what input sequence was being processed. Since there will be a majority of either 1's or 0's in *any* input sequence, we just need to be able to deduce which situation applies as each output bit is being computed. Then, when the majority input-bit type is 0 we use the system already described. When there's a majority of input 1's, however, the above-described process just needs a simple alteration: we then employ a bias of Σh_k and we only perform an add/subtract operation for those tap weights corresponding to a *zero* input bit. Clearly, this approach requires that only addition operations be employed when the bias is $-\Sigma h_k$ and only subtraction operations when the bias is Σh_k. In all cases, we're performing the add/subtract operation for at most *N*/2 tap weights. A variation on this system design which sometimes might be more attractive is to always employ a $-\Sigma h_k$ bias and always perform *additions*, no matter what the majority input data type, but to perform a negation on the final accumulated output value whenever the input-bit-type majority is 1. (Depending on accuracy requirements, a simple conditional bit-inversion—i.e., ones' complement negation—might suffice.)

Now that we can conceive of building a system wherein we know that we must perform at most *N*/2 add/subtract operations in the computation of *each* output data sample, it's evident that another filter design advantage might ensue: if the maximum speed of the adders being employed is a limiting factor in determining the maximum length of the pulse-shaping filter—for a given input data rate—this computational savings can be recognized as a means of saving processing time, thereby permitting the use of longer filters (or, as a means of permitting a system that would otherwise be designed to run at higher data rates).

Rather than employing a biased summation, an alternative implementation would consist of pairing-up tap weights—storing, say, $h_0 + h_1$ and $h_0 - h_1$ rather than h_0 and h_1. Then, a single add/subtract operation of one of these two values would suffice for both taps' contributions to the filter's output value. While this approach would also halve the number of add/subtract operations, while not increasing data storage requirements and not requiring circuitry for determining the majority bit type, it would, however, require bit-type analysis circuitry on each pair of taps and it would not seem to provide the option we've just mentioned of never requiring subtractions. Furthermore, it would always require approximately *N*/2 add/subtract operations whereas the method we've outlined requires *at most* *N*/2, depending on the data.

While we've focused on pulse-shaping filters, the same techniques can also be employed in the design of filters that process one-bit data treating an input 0 bit as 0 rather than -1. In this case we design the same kind of system we've described previously, but the tap coefficients are not doubled and the bias value is either 0 or Σh_k, depending on whether the majority input bit type is 0 or 1. When 0's are in the majority at most *N*/2 additions are performed, and with a majority of 1's present it's at most $N/2$ subtractions. Alternatively, by using a bias offset of 0 or $-\Sigma h_k$ and conditionally negating the output value, we can build the filter in a way that always performs at most $N/2$ additions (i.e., no subtractions are required).

Finding the majority input bit type

We're now faced with the task of designing a system to determine, at the time each output bit is computed, whether the majority of input bits are 0s or 1s. A particularly simple way to do this is as follows: let's start the system with, say, all 0 bits at the outputs of the delay chain elements at the top of the filter of Fig. 3. (These are the system's "initial conditions" and no matter how the system is built, or what values they are given, the first $N-1$ data output values, but only those values, would be affected by this choice.) Let the system contain a *counter* capable of counting from $-N/2$ to $N/2$ if N is even, and from $-(N+1)/2$ to $(N-1)/2$ if N is odd, and let its value be initialized at $-N/2$ or $-(N+1)/2$, respectively. Then, as each new input data bit is encountered and all the previous data bits are shifted across the top delay chain, resulting in a bit being "lost" at the end, we just need to update the counter. When two *zero* bits occur as the new bit and the lost bit, or when two *one* bits occur, we need not update the counter, as no change will have occurred in the majority of input bits present. When a new 1 bit comes in and a 0 bit is lost, however, we then need to increment the counter. Similarly, a new 0 and a lost 1 requires counter decrementing. After each counter update, the counter's sign bit tells us whether the majority of the input data currently being processed is 1 or 0. This type of system is illustrated in the block diagram of Fig. 4.

BPSK and QPSK

The system described above is appropriate for a binary phase-shift keying (BPSK) modulation system. For quadrature phase-shift keying (QPSK), as depicted in Fig. 5, we just build two such pulse-shaping filters, one for each channel.

QAM

Another increasingly popular modulation scheme is quadrature amplitude modulation (QAM), which can be implemented in various degrees of sophistication. Fig. 6 illustrates "16-QAM," which takes four data bits at a time from the incoming bit-stream, i.e., processes four-bit "symbols" and forms each four-bit symbol into two *bit-pairs*, sending one bit-pair onto each of the I and Q channels. Now, two-bit data must be handled. Fig. 7 shows that each channel's pulse-shaping filter (and the subsequent fixed digital and analog processing) must produce one of four possible responses, depending on the particular bit-pair being processed. These responses require scaled copies of the type of pulse-shaping filter response we've already discussed. The constellation diagram of Fig. 6 shows how the four amplitude modulation levels for each channel can be associated with the 16 equally-spaced points (16 symbols) in the I-Q plane. A very straightforward approach to implementing these pulse-shaping filters would be to build a simple FIR filter employing tap multipliers that accept signed three-bit input data (which conveniently encodes the values ±1/4 and ±3/4). While conceptually simple, this would be somewhat wasteful since *two*-bit data on each channel should be employable as each channel's symbols have only four possibilities at each discrete time point. It would also be desirable to employ multiplierless filters, as we did previously. Here's how:

In a manner similar to QPSK, where each bit in a channel's bit-stream uses the 0 state to represent the negative of the 1 state, we accommodate each of the two bits of each 16-QAM channel. For example, if *ab* denotes a channel's bit-pair, we can define these bit values as:

$$a = \begin{cases} 0 \text{ means } -1/2 \\ 1 \text{ means } 1/2 \end{cases} \quad \text{and} \quad b = \begin{cases} 0 \text{ means } -1/4 \\ 1 \text{ means } 1/4 \end{cases}$$

Then, for the bit-pair *ab* we have:

a	*b*	
1	1	$\Rightarrow 3/4$
1	0	$\Rightarrow 1/4$
0	1	$\Rightarrow -1/4$
0	0	$\Rightarrow -3/4$

(This encoding of the bit-pairs *ab* is for the filter designer's convenience. Simple logic circuitry could identify any desired bit-pair patterns with the table's patterns for $\pm 1/4$ and $\pm 3/4$.) Using this encoding, we simply implement two multiplierless FIR filters as previously described, with one filter accepting the one-bit *a*-data and the other accepting the one-bit *b*-data. The *a*-data tap-coefficients are modified from their normal values as follows: first, we use $(1/2)h_k$ for the coefficients h_k of the *a*-data filter, and $(1/4)h_k$ for the coefficients h_k of the *b*-data filter. With the understanding that these are the values of the two filters' coefficients, in terms of the actual filter-tap values, but continuing to refer to the coefficients of these two filters simply as h_k, the same manipulations, i.e., using $-\Sigma h_k$ and $2h_k$, etc., can be used to obtain low-power implementations of the 16-QAM filters. One adder is employed to combine the output bit-streams of the *a*-data and the *b*-data filters, as shown in Fig.8. Obviously, the same principles can be applied to other QAM systems. For 64-QAM, for example, we just use three separate multiplierless filters on each channel: an *a*-data filter, a *b*-data filter, and a *c*-data filter, where the filter coefficients are scaled by 1/2, 1/4, and 1/8, respectively.

A variation on this system design that might be advantageous is to keep all channels' coefficients the same—i.e., let the *a*-channel, *b*-channel, etc., all have the original h_k tap weights, and let all channels process the input data as described for the simple BPSK case. Then, when the channels' outputs are added together, it would be a simple matter to employ hard-wired shifts to account for the different weightings (e.g., 1/2, 1/4,...) on the different channels.

Linear-phase filters

If a filter processing a stream of one-bit data (with, as above, 0 and 1 input bits interpreted as ± 1) has linear phase—illustrated by the even symmetry and even number of tap coefficients in Fig.9 (but, it's clear that any linear-phase symmetry can also be accommodated through simple changes in the details)—there's no obvious way to save on computations by exploiting the linear-phase feature. If the filter were to actually employ multiplication operations, it's well known that about half the multiplies could be saved by reorganizing the way in which the additions are performed. But there are no multiplication operations to save when processing one-bit data—so it seems we're required to perform $N-1$ additions per output sample, just as would be required without linear phase. We've shown in the above discussion that it's possible to reduce these addition operations to at most $N/2$, but this doesn't in any way exploit the filter's linear-phase feature. The question arising from these observations is: isn't it possible to gain some further advantage from the fact that a linear-phase filter has a special coefficient pattern? The answer is *yes*. More specifically, we'll now show that it's possible to reduce the required number of addition/subtraction operations in such a filter to at most $N/3$.

Let's consider the possible data-value combinations for the pair of bits associated at some point in time with equal-coefficient taps in the linear-phase filter of Fig. 9. There are three different possibilities: a 0 and a 1, in which case the taps' contributions to the output sum cancel one another; a pair of 1's, in which case we require that twice the common tap value be added to the output sum; and a pair of 0's, in which case we require that twice the common tap value be subtracted from the output sum. In the spirit of the preceding discussion, we can employ biasing offsets to avoid actually performing some of the add/subtract operations. We will want to avoid the computations associated with the largest of the three sets of possible kinds of tap-pairs to achieve the greatest savings. Thus, for $k = 0, \ldots, (N/2) - 1$, we need to identify the set to which each tap-pair belongs, as follows:

Let $A_k = x(k) \oplus x(N-1-k)$ ($\Rightarrow A_k = 1$ if and only if one bit is 0 and one is 1).

Let $B_k = x(k) \cap x(N-1-k)$ ($\Rightarrow B_k = 1$ if and only if both bits are 1).

Let $C_k = \text{not}\,(x(k) \cup x(N-1-k))$ ($\Rightarrow C_k = 1$ if and only if both bits are 0).

Since, for each k, exactly one of A_k, B_k, C_k is 1, we can build circuitry for computing only, say, A_k and B_k. Such circuitry requires just one half-adder, as shown in Fig. 10. Thus, for N even (although we can also treat the N odd case) we require $N/2$ half-adders whose outputs can be used to identify the optimal bias value and the corresponding tap-coefficient modifications that should be used in evaluating the filter's output value. There are three possible cases, as shown in the following table:

Case	Bias offset	when $A_k = 1$	when $B_k = 1$	when $A_k = B_k = 0$
Case A: (at least 1/3 of the $N/2$ tap-pairs are 0 & 1)	0	—	$2h_k$	$-2h_k$
Case B: (at least 1/3 of the $N/2$ tap-pairs are 1 & 1)	$2\Sigma h_k$	$-2h_k$	—	$-4h_k$
Case C: (at least 1/3 of the $N/2$ tap-pairs are 0 & 0)	$-2\Sigma h_k$	$2h_k$	$4h_k$	—

By employing the computations indicated in the appropriate row of the above table we can eliminate at least 1/3 of the $N/2$ tap-pair add/subtract operations, i.e. we must perform at most $(2/3) \times (N/2) = N/3$ add/subtract operations. When computing the $2N/3$ savings, rounding works in our favor! For example, when building a 20-tap linear-phase filter, for any distribution of input data bits across the 20 taps, there would always be (at least) *four* of the ten tap-pairs that qualify for one of the three cases (A, B, C, in the above table). Thus, we need never process more than six add/subtract operations—a 14/20 = 70% savings in add/subtracts.

In addition to the $N/2$ half-adder circuits, it'll be necessary to build circuitry that detects that at least 1/3 of the half-adder sum or carry outputs are high. This is illustrated in the block diagram of Fig. 11. Depending on the desired level of complexity in circuit design, the *optimal* choice between the A or B outputs (should they *both* be high for some set of input bits) might be made by comparing the actual number of high inputs and selecting the case with the largest number. Alternatively, we might just make an arbitrary choice—say, always pick A. In addition to the circuitry shown in Fig. 11, the individual S_k and C_k bits are also used to select the specific type of add/subtract operation needed for each tap-pair, according to the above table.

General FIR filters

The above techniques can also be used to design completely general FIR digital filters, i.e., filters having B-bit two's-complement input data and M-bit tap-coefficients, with B and M arbitrary positive integers. Our previous discussion concerning QAM provides a structure for filters employing multi-bit input data that's suitable for this application, but it must be modified for correct processing of the input-data sign bits. Let's consider, for example, the input data to be scaled such that it has signed fractional values, i.e., let the B-bit input data have the form $b_0.b_1b_2...b_{(B-1)}$, where b_0 is the sign-bit (having weight -1), b_1 has weight 1/2, b_2 has weight 1/4, etc. Thus, referring to each one-bit-data filter by the subscript number of its data bit, the tap weights for the bit-n filter (for $n = 0, 1, \ldots, B-1$) should be scaled from their original h_k values to become $2^{-n}h_k$ and the corresponding bias offset value for that one-bit-data filter should be, depending on whether the majority of the filter's input bits are 0 or 1, respectively, either 0 or $2^{-n}\Sigma h_k$ for $n \neq 0$ and either 0 or $-\Sigma h_k$ for $n = 0$ (the sign bit).

Setting aside linear-phase issues, an N-tap filter with B-bit data (for N even) employs at most $(B \times N/2) + B - 1$ additions for each output data sample of the Fig. 12 system—a cost that can be reduced at the expense of increased data storage and retrieval circuitry by combining the B bias additions into fewer additions (just one, in the extreme) since all offsets are ultimately accumulated into a common final value. A standard Booth-encoding implementation (with B even) requires $N \times [(B/2) - 1] + N - 1 = (NB/2) - 1$ additions. While both totals are close to $NB/2$, ours is always higher, only exceeding the Booth total by one add when using just a single combined bias value. (Recall, however, unlike Booth, our total is an upper bound; the actual number of adds is data dependent.) In an implementation employing parallel processing, since different combinations of additions must be performed first (in parallel) before other additions can proceed, the specific sizes of N and B could well be the factor that determines which scheme takes less time overall. With both N and B even, the relation that needs to be satisfied for our approach to yield a faster overall system than the Booth-encoding approach is $(N/2) + (B-1) < [(B/2)-1] + (N-1)$, which is equivalent to $N > B+2$.

Turning now to the case of general *linear-phase* FIR filters, the number of adds required for the Fig. 12 system is (at most) $[B \times (N/3)] + (B-1)$ while a Booth-encoded system requires a $\{[(B/2) - 1] \times N/2\} + (N-1)$ total. It follows that our system requires fewer additions when $B < 6$ and N is sufficiently large, namely $N > 12B/(6-B)$. Again, the comparison becomes even more favorable to our approach when fewer than B (combined) bias offset values are employed. In any event, if our concern is speed, and parallel processing is employed, then our method is faster than the Booth-encoding approach when $N/3 + (B-1) < 1 + (B/2 - 1) + (N/2 - 1)$ which can be reduced to the simple relation $N > 3B$.

Transpose FIR filter configurations

We can also employ our techniques with transposed configuration filters. Fig. 13 shows the use of $2h_k$ taps and a $\pm\Sigma h_k$ bias. Here an appropriate stream of $\pm\Sigma h_k$ values must enter the delay chain at the bottom of Fig. 13 to take full advantage of options to avoid additions as they arise. This is easily arranged by storing the bit stream in a shift register prior to its filter entry and making the "majority 0/1" decision as before, but now operating on the shift register contents. A circular read/write storage arrangement (Fig. 14) can be used as an alternative.

Heterogeneous Digit-Serial Finite Field Multipliers and Low-Energy Reed-Solomon Codecs [1]

Leilei Song[2], Keshab K. Parhi[2], Ichiro Kuroda[3], Takao Nishitani[3]
[2] Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN 55455, USA
[3]Digital Signal Processing Lab, NEC Corporation
4-1-1 Miyazaki, Miyamae-Ku, Kawasaki 216, Japan
Email: {*llsong, parhi*}*@ece.umn.edu,* {kuroda, nishitani}@dsp.cl.nec.co.jp

Abstract

Reed-Solomon (RS) codecs are used for error control coding in many applications such as digital audio, digital TV, software radio, CD players, and wireless and satellite communications. Traditionally RS codecs have been implemented using dedicated hardware. Using current scaled technologies, most video and data communications systems can be realized using domain-specific programmable digital signal processors (DS-PDSP) optimized for targeted applications. This paper considers software-based implementation of RS codecs where special instructions are assumed to be used to program finite field multiplication datapaths inside a programmable DSP. To achieve energy reduction at both component level (for one finite field multiplication) and system level (for RS codecs), a novel *heterogeneous digit-serial* approach is proposed. In this digit-serial architecture, the *heterogeneity* corresponds to the use of different digit-sizes in the multiply-accumulate (MAC for polynomial multiplication) and degree reduction (DEGRED for polynomial modulo operation) subarrays. The digit-size parameters for these subarrays are selected so as to minimize energy consumption. The component level analysis leads to the conclusion that the heterogeneous digit-serial architectures lead to the least energy and area consumption. The number of instruction cycles needed to program RS encoding and decoding algorithms using the proposed digit-serial approaches and datapath architectures are then obtained and the total energy consumption, as well as the energy-delay products of these codecs using various datapaths are compared. It is concluded that, for a field order of 8, a parallel MAC unit (of digit-size 8) and a DEGRED unit with digit-size 4 or 2 is the best datapath, with respect to least energy consumption and energy-delay products, for 2-error-correcting RS codec implementations.

1 Introduction

The arithmetic over finite fields are the underlying operations in many error-control coding algorithms. Sophisticated error correcting codes, such as Reed-Solomon codes, are extremely important in both broadband and narrowband digital communication systems. There has been considerable interests in designing dedicated circuits for RS encoders and decoders [1], [2], [3], [4], [5], [6], [7], [8], [9], *etc.* However, designing universal RS codecs which can operate over different finite fields with variable code rate has practical importance and still remains an active research topic [10]. Dramatic increase in the area and hardware complexity with increment in the block length and underlying field order m makes it very difficult to realize universal RS codecs in hardware. Implementing them in software using DSP processors becomes an attractive alternative for designing universal RS codecs. However, due to the lack of hardware support for finite field multiplication,

---

[1] This research was supported in parts by the National Science Foundation under grant number INT-9600372 and the Army Research Office under grant number DA/DAAH-94-G-0405.

sophisticated RS codecs are usually slow in software. A software based RS(16,12) coder was presented in [11], in which general purpose DSP processors or microprocessors were used for programming. If finite field arithmetic would be implemented in a programmable DSP datapath, RS codecs with variable code rate and based on different finite fields could be easily constructed in software. The finite field datapath which is programmable with respect to both field orders and primitive polynomials offers the most flexibility for variable rate codes; the processor based software approach reduces the time-to-market and design cost of RS codecs.

Recently, design for low power has become increasingly important in applications ranging from wireless communications where battery life is the critical factor, to high performance computing where reliability and cooling issues are of important concern. Power reduction techniques have been proposed at all levels of the design hierarchy, from algorithmic and architectural to circuits and technological innovations. In this paper, we use a *hardware software co-design* methodology and achieve power reduction at algorithmic and architectural level. A *domain-specific* DSP [12] is assumed, whose datapath is specialized for finite field operations. The proposed finite field datapath is designed with RS coding applications in mind and can be programmed to realize low-energy RS codecs; the RS codecs are then implemented in software making use of the specialized datapath and instructions to achieve system level energy reduction.

In the finite field $GF(2^m)$, addition and subtraction are bit-independent and can be computed using array of XOR gates. Inversion (as well as division) can be computed iteratively using multiplication. Therefore, to design DSP datapath for finite field arithmetic, we concentrate only on the multiplication. There are two major operations involved in finite field multiplication, namely polynomial multiplication and polynomial modulo operation over $GF(2)$. In a fully parallel multiplier, the polynomial multiplication is implemented using binary-tree of XOR gates; the polynomial modulo operation can only be implemented using linear array structure to accommodate variable primitive polynomials. Linear array structure consumes more energy than tree-type structure due to propagation of switching activities. As a result, while these two operations require similar amount of hardware units, polynomial multiplication consumes about one-fourth of the total energy for one multiplier and polynomial modulo operation consumes about three-fourth of the total energy. As we know, vector-vector multiplication is the computation common to many DSP applications, including RS encoding and decoding. For finite field vector-vector multiplication, polynomial modulo operation can be delayed till the last step, and the intermediate result are obtained by polynomial multiplication and accumulation only. For example, consider vector-vector multiplication over $GF(2^m)$,

$$\begin{bmatrix} A_0 A_1 \cdots A_{n-1} \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ \cdots \\ B_{n-1} \end{bmatrix} = A_0B_0 + A_1B_1 + \cdots + A_{n-1}B_{n-1}, \tag{1}$$

where A_i and B_i, $0 \leq i \leq n-1$, are elements of $GF(2^m)$ with wordlength equal to m. The intermediate products, A_iB_i, have wordlength $2m-1$ and are obtained using polynomial multiplications only. Then after accumulating these intermediate results, only one polynomial modulo operation is required to reduce the degree of the final result from $2m-2$ to $m-1$. The saving in energy can be roughly estimated as follows:

$$\frac{pow \times n - (\frac{1}{4}pow \times n + \frac{3}{4}pow)}{pow \times n} = \frac{\frac{3}{4}(n-1)}{n} \approx 75\%, \tag{2}$$

where *pow* denotes the average power consumption of one parallel multiplier. In this way, the vector-vector multiplication is performed with only 25% energy consumption for arithmetic operations, at the expense of increasing instruction count by 1. Based on this observation, we propose the idea of having two separate subarrays in finite field datapath and two separate instructions for polynomial multiplication and polynomial modulo operations. We also propose use of *heterogeneous digit-serial* architectures, where different digit-sizes are assigned to different multiplication subarrays. The performance of this heterogeneous digit-serial approach is evaluated and compared with that using one parallel multiplier in the programmable DSP datapath for one finite field multiplication operation. This is referred to as *component level* performance analysis. The digit-serial datapath architectures are then used to design energy efficient RS codecs, where performance characteristics are analyzed based on area, total energy consumption and instruction count for RS encoding and decoding. This is referred to as *system level* performance analysis.

This paper is organized as follows. Section 2 gives a brief overview of some basic concepts, algebraic structures of finite field. Section 3 presents the heterogeneous digit-serial approach. Its performance characteristics for realizing one finite field multiplication operation are evaluated and compared with other design alternatives. In section 4, this heterogeneous digit-serial approach is used to design datapath architecture of a DS-PDSP processor targeted for low-energy RS codecs applications.

2 Mathematical Background

Knowledge of basic finite field concepts and properties is assumed, as covered in [13], [14], [15]. Here we briefly review some concepts and properties related to arithmetic operations. In what follows, the symbols in $GF(2^m)$, and the bits of symbols (elements of $GF(2)$), are represented using upper and lower case variables, respectively. Unless specified explicitly, "+" and "·" denote logic XOR and AND operations, respectively.

Finite field $GF(2^m)$ contains 2^m elements. It is an extension field of $GF(2)$, which has elements 0 and 1. All finite fields contain a zero element, a unit element, a primitive element and have at least one primitive polynomial $p(x) = x^m + p_{m-1}x^{m-1} + \cdots + p_1 x + p_0$ over $GF(2)$ associated with it. The primitive element α generates all nonzero elements of $GF(2^m)$ and is a root of the primitive polynomial $p(x)$. Therefore, we have

$$GF(2^m) = \{0, \alpha^0, \alpha^1, \cdots, \alpha^{2^m-2}\}, \tag{3}$$

which is referred to as *exponential representation*, and

$$\alpha^m = p_{m-1}\alpha^{m-1} + p_{m-2}\alpha^{m-2} + \cdots + p_1\alpha + p_0. \tag{4}$$

Using (4), all elements of $GF(2^m)$ can also be represented as polynomials of α with a degree less than m by performing a *mod* $p(\alpha)$ operation to $\alpha^t, 0 \leq t \leq 2^m - 2$. Hence, we have

$$\begin{aligned} GF(2^m) &= \{A | A = a_{m-1}\alpha^{m-1} + a_{m-2}\alpha^{m-2} + \cdots + a_1\alpha + a_0, \\ &\quad \textit{where } a_j \in GF(2),\ 0 \leq j \leq m-1\}, \end{aligned} \tag{5}$$

which is referred to as *polynomial representation.* The polynomial representation is generally used for finite field arithmetic operation, where addition is carried out using bit-independent XOR operations, and multiplication is carried out using polynomial multiplication and modulo operations. The polynomial representation is based on the *basis representation*, where $\{1, \alpha, \alpha^2, \cdots, \alpha^{m-1}\}$ is known as the *standard basis* which is also referred to as the polynomial basis or conventional basis.

3 Heterogeneous Digit-Serial Multipliers

Digit-serial architectures are best suited for systems requiring moderate sample rate and where area and power consumption are critical [16]. In this section, we propose a *heterogeneous* digit-serial approach for the design of finite field datapath in the DS-PDSP processor. In this approach, one parallel finite field multiplication is decomposed into *MAC* and *DEGRED* subarrays. Digit-sizes $D1$ and $D2$ are assigned to *MAC*

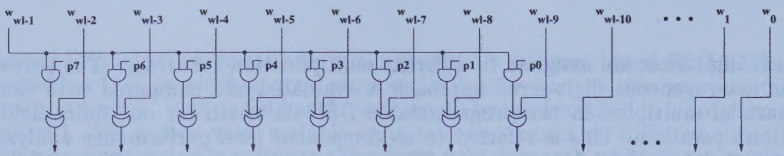


Figure 1: Circuit for One row of $DEGRED1$ array

and $DEGRED$ arrays, respectively, and multiplication over $GF(2^m)$ is performed digit-serially in software. It is essentially a digit-serial/parallel approach, where multiplicand bits are processed in parallel, while multiplier bits are divided into digits and processed digit-serially. This approach is intended to offer more programming flexibilities and achieve system level energy reduction for real applications implemented using this DSP datapath architecture.

Since most pratical RS codecs are operated over $GF(2^8)$, we only consider finite field multiplication over $GF(2^8)$ in this section. The corresponding datapath architectures are built for $GF(2^8)$. However, they can be programmed to compute multiplications over finite field $GF(2^m)$, for $m \leq 8$, by applying zero-padding and extra shifting operations to the operands.

3.1 Heterogeneous digit-serial finite field multipliers

In the heterogeneous digit-serial approach, two subarrays with different digit-sizes, one $MAC\ 8 \times D1$ array of digit-size $D1$ and one $DEGRED\_D2$ array with digit-size $D2$ are assumed in the DSP datapath, and two corresponding instructions, MAC and $DEGRED$, are assigned to them, which can be used to program finite field multiplications. $MAC8 \times D1$ array computes the product of two polynomials over $GF(2)$, one of degree 7 and the other of degree $D1 - 1$, and accumulates their product to some initial value or some intermediate result from previous computation. $DEGRED\_D2$ array performs polynomial modulo operation and reduces the degree of input polynomial by $D2$. It consists of $D2$ rows of $DEGRED1$, whose structure is as shown in Figure 1, where wl denotes the input wordlength.

Various digit-serial multipliers can be obtained by varying $D1$ and $D2$. To simplify our analysis, we assume that $D1$ can only be divisors of $M = 8$, i.e., 8, 4, 2, 1; $D2$ can only be divisors of $D1$. Hence, there are totally 10 possible combinations of hardware units:

$MAC\ m \times D1$	$DEGRED\_D2$
$MAC8 \times 8$	$DEGRED7$
$MAC8 \times 8$	$DEGRED4$
$MAC8 \times 8$	$DEGRED2$
$MAC8 \times 8$	$DEGRED1$
$MAC8 \times 4$	$DEGRED4$
$MAC8 \times 4$	$DEGRED2$
$MAC8 \times 4$	$DEGRED1$
$MAC8 \times 2$	$DEGRED2$
$MAC8 \times 2$	$DEGRED1$
$MAC8 \times 1$	$DEGRED1$

Note that when $D1 = M = 8$, the maximum degree of the intermediate result is equal to 15 which requires degree reduction by 7 only. Hence instead of having $D2 = D1 = 8$, we set maximum value of $D2$ to 7.

Table 1: Performance comparison of finite field multipliers

Multiplier	**Crit.** **path**	**Latency** # of cycles	**Area** # of trans.	**Energy** *pJ*
1. Semi-sys(non-pipe)	16D	1	1720	16.402
2. Semi-sys(one-pip)	8D	2	2008	10.174
3. Pal-Mult(non-pipe)	19D	1	1776	12.256
4. Pal-mult(one-pipe)	10D	2	2028	8.986
5. $MAC8 \times 8$+$DEGRED7$	14D	2	1956	10.632
6. $MAC8 \times 8$+$DEGRED4$	8D	3	1620	8.99
7. $MAC8 \times 8$+$DEGRED2$	5D	5	1396	7.678
8. $MAC8 \times 8$+$DEGRED1$	5D	8	1284	7.443
9. $MAC8 \times 4$+$DEGRED4$	8D	3	1124	9.244
10. $MAC8 \times 4$+$DEGRED2$	4D	5	900	7.932
11. $MAC8 \times 4$+$DEGRED1$	4D	8	788	7.697
12. $MAC8 \times 2$+$DEGRED2$	4D	5	652	8.588
13. $MAC8 \times 2$+$DEGRED1$	3D	8	540	8.353
14. $MAC8 \times 1$+$DEGRED1$	2D	8	416	9.913

3.2 Comparison with other design alternatives

In this subsection, the performance characteristics of the digit-serial multipliers are compared with two parallel multipliers, the lsb-first (least-significant-bit-first) semi-systolic array multiplier presented in [17] and the fully parallel multiplier, where the two subarrays, $MAC8 \times 8$ and $DEGRED7$, are combined into one unit. The parallel multiplier in DSP datapath is treated as one unit and one instruction is assigned to it, which includes both polynomial multiplication and modulo operations.

Performance comparison of different types of finite field multipliers are summarized in Table 1, in which critical path computation time is expressed in terms of total number of gate delays; latency is defined as the total number of clock cycles required for one multiplication; area is provided in terms of total number of transistors; the total energy consumption for one multiplication operation using various multipliers are estimated using the HEAT tool proposed in [18] and the results are obtained at a clock frequency of 50MHZ, using 0.35μ technology, with 3.3 *Volts* supply voltage. Both non-pipelined (denoted as non-pipe) and one-level pipelined (denoted as one-pipe) parallel multipliers are considered.

For digit-serial multipliers, area decreases as digit-sizes $D1$ and $D2$ decrease. Although digit-serial multipliers have longer latency, they generally have smaller critical path and can be operated at higher clock rate. The total energy for one multiplication operation does not change with the operating clock frequency (recall that energy consumption in a digital CMOS circuit can be approximated as $E = \alpha CV^2$, where α is the activity factor). Therefore, one digit-serial multiplication generally has less area, consumes less energy while takes about the same total computation time as parallel multipliers. Nevertheless, the ideas of having two subarrays in the DSP datapath and digit-serial algorithms are not intended to improve the performance of one multiplication operation, but to improve the overall performance of all the arithmetic operations in a real application program. The programming flexibility offered by this heterogeneous digit-serial approach can be utilized to reduce system level energy consumption, as illustrated in the next section by designing software based low-energy RS codecs.

Table 2: Instruction count for RS(n,k) coding with $n-k=4$ and over $GF(2^8)$ with one parallel multiplier DSP datapath (table on the left) or with two separate arrays, $MAC8\times 8$ and $DEGRED7$ (table on the right)

Instr.	Encoding	Decoding
MULT8×8	$4n-16$	$6n{+}37$
MEM	$5n-16$	$6n{+}21$
MIS	4	$4n{+}41$
Total	$9n$-28	$16n{+}99$

Instr.	Encoding	Decoding
$MAC8\times 8$	$4n-16$	$5n{+}37$
$DEGRED7$	4	$2n{+}42$
MEM	$5n-16$	$6n{+}21$
MIS	4	$4n{+}41$
Total	$9n$-24	$17n{+}141$

4 Software-Based Low-Energy RS Codecs

In this section, a simplified DS-PDSP core model is assumed with simple instruction set and limited number of data registers. Datapath performance for 2-error-correcting RS(n,k) codecs using generator matrix based encoding and Peterson-Gorenstein-Zierler decoding algorithms are then analyzed and compared. To simplify the comparison, we compare the datapath containing one parallel multiplier with those having two separate MAC and $DEGRED$ subarrays of different digit-sizes.

4.1 The assumed DS-PDSP model

A domain-specific programmable DSP processor (DS-PDSP) is assumed whose datapath is designed to achieve low-energy and high-performance RS encoder and decoder. A simple processor architecture is assumed, with one instruction memory, one data memory, one data bus and one main bus. A simple instruction set is assumed, which contains less than 16 instructions, including multiply (one instruction is required if the datapath has one parallel multiplier; two instructions are required if the datapath contains separate MAC and $DEGRED$ arrays), Load/Store, logic and shifting, and branch instructions.

4.2 Low-energy 2-error-correcting RS(n,k) codecs over $GF(2^8)$

In this subsection, low-energy 2-error-correcting RS(n,k) codecs are considered. Generator-matrix based RS encoding method is used which requires RAM with storage size $4k$ bytes for precomputed-stored generator matrix. Peterson-Gorenstein-Zierler algorithm is used for RS decoding. Syndrome computation is based on precomputed lookup table which requires RAM with storage size of $4n$ bytes. The reader may refer to [13] for a detailed explanation of RS encoding and decoding algorithms.

We first illustrate the idea for low-energy RS codecs by comparing the performance of datapath with one parallel multiplier and that with two separate subarrays, $MAC8$ and $DEGRED7$. Heterogeneous digit-serial approach are then applied to achieve the best trade-off between energy consumption and performance.

4.2.1 Energy reduction approaches

Suppose that one parallel finite field multiplier is available in a DSP datapath. Instruction count for RS encoding and decoding programs can be computed as shown in the left table in Table 2, where **MULT**8×8 denotes total number of parallel multiply-accumulation instructions; **MEM** denotes total number of memory access instructions, including both LOAD and STORE; **MIS** denotes total number of other miscellaneous instructions. The energy for arithmetic operations can be computed as:

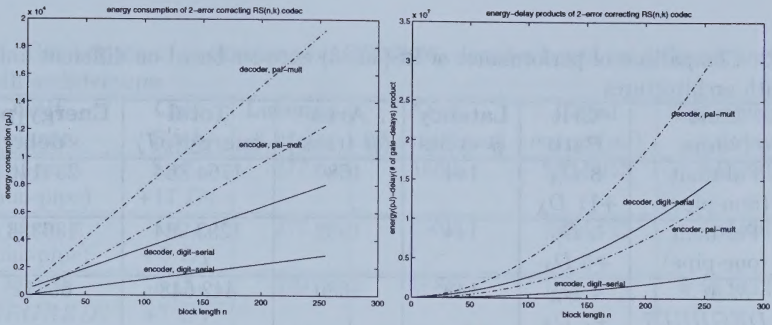


Figure 2: Energy and energy-delay product plots for 2-error-correcting RS codecs over $GF(2^8)$ using different DSP datapaths

$$\begin{aligned} encoder: &\quad 12.256 \times (4n - 16) = 49.024n - 196.096(pJ) \\ decoder: &\quad 12.256 \times (6n + 37) = 73.536n + 453.472(pJ). \end{aligned} \tag{6}$$

Suppose that two subarrays, $MAC8 \times 8$ and $DEGRED7$, are available in DSP datapath. In this case, polynomial modulo operation $DEGRED7$ need not be performed for each finite field multiplication. This is especially true for vector-matrix multiplication, where all intermediate results have wordlength $2m - 1$ and are computed using MAC instruction only. $DEGRED7$ is required at the end to reduce the degree of the final result from $2m - 2$ to $m - 1$. Notice that $DEGRED7$ subarray consumes about 3 times energy as $MAC8 \times 8$ subarray does. Hence if the program can be written in such a way which minimizes the total number of $DEGRED7$ instructions without significantly increasing the total number of arithmetic instructions, then the total energy consumption for arithmetic operations in RS codecs can be reduced dramatically. The RS coding algorithms are written following this rule and the instruction count is computed as shown in the right table in Table 2. The total energy consumption for arithmetic operations can be computed as:

$$\begin{aligned} encoder: &\quad 2.858 \times (4n - 16) + 7.774 \times 4 = 11.432n - 14.632(pJ) \\ decoder: &\quad 2.858 \times (5n + 37) + 7.774 \times (2n + 42) = 29.838n + 432.254(pJ). \end{aligned} \tag{7}$$

The energy consumption and the energy-delay products for 2-error-correcting RS codecs based on the above two datapaths are plotted in Figure 2, from which we conclude that using datapath with two subarrays, about $2/3 - 3/4$ energy reduction for RS encoders and $3/5$ energy reduction for RS decoders can be achieved; the energy-delay products for both encoder and decoder have also been reduced, which implies that the proposed datapath architecture achieves the best trade-off between energy reduction and performance improvement.

4.2.2 Heterogeneous digit-serial approach for low-energy 2-error-correcting RS(40,36) codecs over $GF(2^8)$

In this subsection, the heterogeneous approach is used to design RS codecs, and total energy consumption or the energy-delay product values are used to select the best digit-size combinations $(D1, D2)$.

RS(40,36) code over $GF(2^8)$ is a shortened RS code. The performance of RS encoder is shown in Table 3. Notice that since the MAC and $DEGRED$ can be operated in parallel, the latency of arithmetic operations usually is less than the sum of total numbers of MAC and $DEGRED$ operations.

Table 3: Comparison of performance of RS(40,36) encoder based on different finite field datapath architectures

Unit Name	**Crit Path**	**Latency** # cycles	**Area** (# trans.)	**Total energy**(pJ)	**Energy(pJ)** ×**delay**
Pal-mult (non-pipe)	$8\ D_A$ $+11\ D_X$	144	1680	1764.864	254140
Pal-mult (one-pipe)	$5\ D_A$ $+5\ D_X$	144	1932	1293.984	186333
$MAC8$ $+DEGRED7$	$7\ D_A$ $+7\ D_X$	148	1680	442.648	65512
$MAC8$ $+DEGRED4$	$4\ D_A$ $+4\ D_X$	152	1344	436.080	66284
$MAC8$ $+DEGRED2$	$1\ D_A$ $+4\ D_X$	160	1120	430.832	68933
$MAC8$ $+DEGRED1$	$1\ D_A$ $+4\ D_X$	172	1008	429.892	73941
$MAC4$ $+DEGRED4$	$4\ D_A$ $+4\ D_X$	296	896	472.656	139906
$MAC4$ $+DEGRED2$	$1\ D_A$ $+3\ D_X$	304	672	467.408	142092
$MAC4$ $+DEGRED1$	$1\ D_A$ $+3\ D_X$	316	560	466.468	147404
$MAC2$ $+DEGRED2$	$2\ D_A$ $+2\ D_X$	592	448	561.872	332628
$MAC2$ $+DEGRED1$	$1\ D_A$ $+2\ D_X$	604	336	560.932	338803
$MAC1$ $+DEGRED1$	$1\ D_A$ $+1\ D_X$	1180	224	785.572	926975

Similar analysis is performed for RS decoder and the performance characteristics are shown in Table 4.

From Tables 3 and 4, we can conclude that RS codecs based on $MAC8+DEGRED2$, $MAC8+DEGRED1$, $MAC4+DEGRED2$ and $MAC4+DEGRED1$ digit-serial datapaths consume less energy; RS codecs based on $MAC8+DEGRED4$ and $MAC8+DEGRED2$ have the smallest energy-delay product values, hence achieve the best trade-off between energy and performance. We conclude that the datapath with $MAC8+DEGRED2$ is the best design for simultaneous minimization of energy and energy-delay product.

5 Conclusions and Future Work

A hardware-software codesign approach has been used to design the datapath for a domain-specific DSP processor. Special datapath architectures for finite field arithmetic operations have been constructed with low-energy RS codecs application in mind. Energy reduction has been achieved at both component and system levels.

In this paper, a novel heterogeneous digit-serial approach for finite field multiplication was proposed and used to design a programmable DSP datapath. This approach

Table 4: Comparison of performance of RS(40,36) decoder based on different finite field datapath architectures

Unit Name	**Crit Path**	**Latency** # cycles	**Area** (# transistor)	**Total** energy(pJ)	**Energy(pJ)** ×**delay**
Pal-mult (non-pipe)	$8\ D_A$ $+11\ D_X$	277	1680	3394.912	940390
Pal-mult (one-pipe)	$5\ D_A$ $+5\ D_X$	277	1932	2489.122	689486
$MAC8$ $+DEGRED7$	$7\ D_A$ $+7\ D_X$	359	1680	1625.774	583653
$MAC8$ $+DEGRED4$	$4\ D_A$ $+4\ D_X$	441	1344	1302.810	574539
$MAC8$ $+DEGRED2$	$1\ D_A$ $+4\ D_X$	605	1120	1120.786	678075
$MAC8$ $+DEGRED1$	$1\ D_A$ $+4\ D_X$	851	1008	1079.516	918668
$MAC4$ $+DEGRED4$	$4\ D_A$ $+4\ D_X$	603	896	1363.008	821894
$MAC4$ $+DEGRED2$	$1\ D_A$ $+3\ D_X$	767	672	1180.984	905815
$MAC4$ $+DEGRED1$	$1\ D_A$ $+3\ D_X$	1013	560	1139.714	1154530
$MAC2$ $+DEGRED2$	$2\ D_A$ $+2\ D_X$	1091	448	1336.456	1458073
$MAC2$ $+DEGRED1$	$1\ D_A$ $+2\ D_X$	1337	336	1295.186	1731664
$MAC1$ $+DEGRED1$	$1\ D_A$ $+1\ D_X$	1985	224	1714.031	3402352

offers programmers the most flexibility by providing separate hardware (hence distinct instructions) and different digit-sizes for polynomial multiplication and modulo operations. Low-energy RS codecs were then constructed in software making use of this special datapath. The analysis shows that the heterogeneous digit-serial approach with two sub-arrays, $MAC8$ and $DEGRED4$, or $MAC8$ and $DEGRED2$, are best suited datapath combinations for RS codecs with respect to energy consumption and performance.

Future work includes analyzing performance of more sophisticated RS decoding algorithms, such as Berlekamp-Massey algorithm, Euclidean algorithm [13] and modified Euclidean algorithm [7] using the heterogeneous digit-serial datapath, for more sophisticated RS codes; and developing supporting software programs for low energy RS codecs.

References

[1] I.S. Hsu, I. S. Reed, T. K. Truong, K. Wang, C.S. Yeh, and L. J. Deutsch, "The VLSI Implementation of a Reed-Solomon Encoder Using Berlekamp's Bit-Serial Multiplier Algorithm", *IEEE Trans. on Computers*, vol. c-33, pp. 906–911, Oct. 1984.

[2] G. Seroussi, "A Systolic Reed-Solomon Encoder", *IEEE Trans. on Information Theory*, vol. 37, pp. 1217–1220, July 1991.

[3] M. A. Hasan and V. K. Bhargava, “Architecture for a Low Complexity Rate-Adaptive Reed-Solomon Encoder”, *IEEE Trans. on Computers*, vol. 44, pp. 938–942, July 1995.

[4] S. K. Jain and K. K. Parhi, “Efficient Standard Basis Reed-Solomon Encoder”, *in Proc. of 1996 IEEE Int. Conf. of Acoustics, Speech, and Signal Processing*, Atlanta, May 1996.

[5] J.L. Politano and D. Depery, “A 30 Mbits/s (255, 223) Reed-Solomon Decoder”, Technical report, Alcatel Atfh - France.

[6] K. Y. Liu, “Architecture for VLSI Design of Reed-Solomon Decoders”, *IEEE Trans. on Computers*, vol. c-33, pp. 178–189, Feb. 1984.

[7] H. M. Shao, T. K. Truong, L. J. Deutsch, J. H. Yuen, and I. S. Reed, “VLSI Design of a Pipeline Reed-Solomon Decoder”, *IEEE Trans. on Computers*, vol. c-34, pp. 393–403, May 1985.

[8] H. M. Shao and I. S. Reed, “On the VLSI Design of a Pipeline Reed-Solomon Decoder Using Systolic Arrays”, *IEEE Trans. on Computers*, vol. 37, pp. 1273–1280, Oct 1988.

[9] J.M. Hsu and C.L. Wang, “An Area-Efficient VLSI Architectures for Decoding of Reed-Solomon Codes”, *in IEEE Proc. of ICASSP-96*, pp. 3291–3294, Atlanta, May 1996.

[10] S. B. Wicker and V. K. Bhargava, *Reed-Solomon Codes and Their Applications*, IEEE Press, New York, NY, 1994.

[11] Y. R. Shayan, T. Le-Ngoc, and V. K. Bhargava, “Design of Reed-Solomon (16,12) Codec for North American Advanced Train Control System”, *IEEE Transitions on Vehicular Technology*, vol. 39, pp. 400–409, November 1990.

[12] I. Verbauwhede and M. Touriguian, “A Low Power DSP Engine for Wireless Communications ”, *in Proc. of VLSI Signal Processing, IX*, pp. 471–480, San Francisco, CA, October 1996.

[13] R. E. Blahut, *Theory and Practice of Error Control Codes*, Addison Wesley, 1984.

[14] W. W. Peterson and E. J. Weldon, *Error-Correcting Codes*, The MIT Press, 1972.

[15] R. J. McEliece, *Finite Fields for Computer Scientists and Engineers*, Kluwer Academic, 1987.

[16] R.I. Hartley and K.K. Parhi, *Digit-Serial Computation*, Kluwer Academic, 1995.

[17] S. K. Jain, L. Song, and K. K. Parhi, “Efficient Semi-Systolic Architectures for Finite Feild Arithmetic”, *IEEE Trans. on VLSI Systems*, vol. , to appear.

[18] J. H. Satyanarayana and K. K. Parhi, “HEAT: Hierarchical Energy Analysis Tool”, *in Proc. 33rd ACM/IEEE Design Automation Conf.*, pp. 9–14, Las Vegas, June 1996.

Reduced Complexity Decision Feedback Equalization for High-Speed Communications over Sparse Channels*

Ian J. Fevrier, Saul B. Gelfand, and Michael P. Fitz†
Department of Electrical and Computer Engineering
Purdue University
W. Lafayette, IN 47907-1285

Abstract

Two modified DFE structures are presented for the equalization of sparse channels typically encountered in high definition television (HDTV) terrestrial broadcast. Unlike the usual DFE, these structures allow the channel's sparseness to be efficiently exploited for equalizer tap allocation before it is degraded by feedforward filtering. The *thresholded complete feedback equalizer* (TCFE) fully exploits channel sparseness. The *thresholded partial feedback equalizer* (TPFE) partially exploits channel sparseness, but has a simpler implementation. Both structures yield large reductions in complexity while maintaining performance comparable to the conventional DFE, hence overcoming a key bottleneck in the equalization of typical HDTV terrestrial channels.

1 Introduction

The high data rates and large delay spreads encountered in the transmission of high definition television (HDTV) signals over terrestrial channels makes high performance equalization computationally expensive [5]. The decision feedback equalizer (DFE) with its relatively low complexity and near optimal performance is potentially a good solution to the problem. However, even the DFE can be overly complex when delay spreads extending over hundreds of symbols are encountered.

Some methods of DFE complexity reduction are described in the literature ([6], [7]), however these methods are not tailored to the HDTV terrestrial channel. These channels are characterized by a relatively short precursor response and a long, sparse postcursor response [5] (see Figure 3). The feedforward filter of the conventional DFE degrades this sparseness property resulting in an increase in the significant feedback filter taps. This suggests that utilization of channel sparseness for efficient cancelation of postcursor ISI should be done prior to precursor equalization.

We first describe a modified DFE structure called the *thresholded partial feedback equalizer* (TPFE) which has a simple implementation and partially cancels the postcursor ISI before feedforward filtering. We then describe a structure called the *thresholded complete feedback equalizer* (TCFE) which has a more complex implementation but completely cancels all postcursor ISI prior to precursor equalization. The advantage of the TPFE and TCFE structures proposed here is complexity reduction (and *not* performance improvement) over the usual DFE applied to sparse channels. Both structures are shown

---

[*] This work was supported by Thomson Consumer Electronics.

[†] Michael P. Fitz is with the Department of Electrical Engineering, The Ohio State University, Columbus, OH 43210-1272

to yield significant reductions in complex multiply operations that must be implemented in hardware for speed (H-CMULTs), while maintaining performance comparable to the conventional DFE on HDTV terrestrial channels.

2 Channel Model

Let $\{I_n\}$ be a T-rate, discrete M-ary sequence of independent and identically distributed symbols to be transmitted across a discrete-multipath channel. The transmission path, which includes transmitter and receiver square-root raised cosine filters, is modelled as a T-spaced (causal) FIR filter $\{c_n : 0 \leq n \leq L\}$. The symbol-rate sampled received signal is thus

$$y_k = y(kT) = \sum_{n=0}^{L} c_n I_{k-n} + n_k. \tag{1}$$

where n_k signifies samples of a complex, zero mean, bandlimited Gaussian process (which is colored due to processing by receiver pulse shaping filter) with variance N_0. The receiver signal to noise ratio is calculated according to

$$SNR = \frac{\sigma_I^2}{N_0} \sum_n |c_n|^2, \tag{2}$$

where $\sigma_I^2 = E\{|I_n|^2\}$.

3 DFE Structure

We first consider a DFE with feedforward filter $F(D) = \sum_{n=-K}^{0} f_n D^n$, and feedback filter[1] $B_d(D) = \sum_{n=1}^{L} b_{d,n} D^n$. The DFE output before slicing is given by

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n y_{k-n} + \sum_{n=1}^{L} b_{d,n} \tilde{I}_{k-n}, \tag{3}$$

where $\widetilde{I}_n$ are the equalizer output decisions. $K + L + 1$ H-CMULTs per output decision are required. The length of the feedback filter ensures that all postcursor ISI is canceled. In practice the feedforward filter length is chosen on the order of one to five times the length of the channel precursor. For HDTV terrestrial channels, the postcursor response, which is much much longer in duration than the precursor response, mainly determines DFE computational complexity.

It is reasonable to subject the DFE feedback filter to a tap-magnitude threshold so as to (indirectly) exploit the channel sparseness in reducing H-CMULTs associated with postcursor ISI cancellation, i.e.

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n y_{k-n} + \sum_{n \in \Omega_d} b_{d,n} \tilde{I}_{k-n}, \tag{4}$$

---

[1] We will subscript by d,p, and c when it is necessary to distinguish between DFE, PFE, and CFE quantities respectively.

where $\Omega_d = \{n : 1 \le n \le L, |b_{d,n}| \ge \eta_d\}$. We call this a *thresholded decision feedback equalizer* (TDFE). The TDFE requires $K + |\Omega_d| + 1$ H-CMULTs. Sparseness of $\{c_n\}$ does not in general imply sparseness of the the combined channel and feedforward filter response $\{f_n\} * \{c_n\}$, and hence $\{b_{d,n}\}$. If the channel sparseness is to be fully exploited in an H-CMULT reduction strategy, it must somehow be utilized before it is degraded or destroyed by the operation of the DFE feedforward filter.

4 TPFE Structure

To more effectively exploit channel sparseness for H-CMULT reduction, we develop equalizers which subtract out postcursor response before feedforward filtering. We proceed by removing some of the postcursor response

$$\tilde{y}_k = y_k - \sum_{n=L_0}^{L} c_n \tilde{I}_{k-n}. \tag{5}$$

$\tilde{y}_k$ is then used as input to a conventional DFE whose pre-decision output is

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n \tilde{y}_{k-n} + \sum_{n=1}^{L_1} b_n \tilde{I}_{k-n}. \tag{6}$$

The evaluation of $\widehat{I}_k$ requires $\tilde{y}_{k+K}$ and thus $\tilde{I}_{k+K-L_0}$. This implies $L_0 > K$, and without loss of generality we choose $L_0 = K + 1$. Assuming correct past decisions, we have

$$\tilde{y}_{k+n} = \sum_{m=0}^{K} c_m I_{k+n-m} + n_{k+n}, \qquad n = 0, ..., K,^2 \tag{7}$$

which shows that a feedback filter of length $L_1 = K$ is sufficient to cancel the residual postcursor ISI in $\widehat{I}_k$. Hence we arrive at the following *partial feedback equalizer* (PFE):

$$\tilde{y}_k = y_k - \sum_{n=K+1}^{L} c_n \tilde{I}_{k-n} \tag{8}$$

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n \tilde{y}_{k-n} + \sum_{n=1}^{K} b_{p,n} \tilde{I}_{k-n}. \tag{9}$$

The implementation of the PFE (See Figure 1) is similar to the DFE in that the inputs to the PFE are generated by a tapped delay line

$$\tilde{y}_{k+K} = y_{k+K} - \sum_{n=K+1}^{L} c_n \tilde{I}_{k+K-n}. \tag{10}$$

$$\tilde{y}_{k+n} = D(\tilde{y}_{k+n+1}), \qquad n = 0, ..., K-1. \tag{11}$$

---

[2] The noise process $\{n_k\}$ should not be confused with index $n = 0, ..., K$.

For sparse channels we propose the *thresholded partial feedback equalizer* (TPFE)

$$\tilde{y}_{k+K} = y_{k+K} - \sum_{n \in \Omega_{p1}} c_n \tilde{I}_{k+K-n}. \tag{12}$$

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n \tilde{y}_{k-n} + \sum_{n \in \Omega_{p2}} b_{p,n} \tilde{I}_{k-n}. \tag{13}$$

$$\Omega_{p1} = \{n : K+1 \leq n \leq L,\ \ |c_n| > \eta_{p1}\}. \tag{14}$$

$$\Omega_{p2} = \{n : 1 \leq n \leq K,\ \ |b_{p,n}| > \eta_{p2}\}. \tag{15}$$

We want to pick commensurate thresholds for the channel and feedback filter taps, i.e., we want to pick η_{p1}, η_{p2} in some compatible way. Consider the increase in MSE associated with deleting a single tap c_n or $b_{p,m}$ according to $|c_n| < \eta_{p1}$, or $|b_{p,m}| < \eta_{p2}$ respectively. If c_n is deleted, the MSE increases by $\sigma_I^2 |c_n|^2 \left\| f \right\|^2$, while if $b_{p,m}$ is deleted the MSE increases by $\sigma_I^2 |b_m|^2$. Hence for these contributions to be equal we require $\eta_{p2} = \eta_{p1} \left\| f \right\|$. Note that certain postcursor channel taps are thresholded prior to feedforward filtering so that the sparseness of $\{c_n\}$ is more effectively exploited with the TPFE relative to the TDFE. The TPFE requires $K + |\Omega_{p1}| + |\Omega_{p2}| + 1$ H-CMULTs per output decision.

5 TCFE Structure

To further exploit channel sparseness in reducing hardware complexity, we develop feedback equalizers which subtract out all postcursor ISI before feedforward equalization. We proceed by removing more of the postcursor response from y_{k+n}, $0 \leq n \leq K$ in a tap-dependent way

$$\tilde{y}_k^{(l)} = y_k - \sum_{n=L_k^{(l)}}^{L} c_n \tilde{I}_{k-n}, \qquad k = l, ..., l+K, \tag{16}$$

and then using these as inputs to a DFE structure

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n \tilde{y}_{k-n}^{(k)} + \sum_{n=1}^{L_1} b_{c,n} \tilde{I}_{k-n}. \tag{17}$$

The parameters $L_k^{(l)}$, $l \leq k \leq l+K$, and L_1 are subject to some constraints. First the evaluation of $\widehat{I}_k$ requires $\tilde{y}_{e,k+n}^{(k)}$, $0 \leq n \leq K$, and thus $\tilde{I}_{k+n-L_{k+n}^{(k)}}$, $0 \leq n \leq K$. This implies $L_{k+n}^{(k)} > n$, $0 \leq n \leq K$, and without loss of generality we choose $L_{k+n}^{(k)} = n+1$, $0 \leq n \leq K$, or equivalently $L_k^{(l)} = k - l + 1$, $l \leq k \leq l+K$. Second, assuming correct past decisions, we have

$$\tilde{y}_{k+n}^{(k)} = \sum_{m=0}^{n} c_m I_{k+n-m} + n_{k+n}, \qquad n = 0, ..., K, \tag{18}$$

which shows that there is no residual postcursor ISI in $\widehat{I}_k$ and therefore no need for a feedback filter ($L_1 = 0$). Hence we arrive at the following *complete feedback equalizer* (CFE)

$$\tilde{y}_k^{(l)} = y_k - \sum_{n=k-l+1}^{L} c_n \tilde{I}_{k-n}, \qquad k = l, ..., l+K, \tag{19}$$

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n \tilde{y}^{(k)}_{k-n}. \tag{20}$$

Note that if we assume correct past decisions, the minimization of the MSE, $E\{|I_k - \widehat{I}_k|^2\}$ yields the same feedforward filters $\{f_n\}$, and consequently the same outputs for the CFE, PFE, and DFE. The implementation of the CFE (see Figure 2) is different and more complex than the PFE and DFE in that the inputs to the CFE are generated by a recursion which is not a simple tapped delay line,

$$\tilde{y}^{(k)}_{k+K} = \tilde{y}_{k+K} = y_{k+K} - \sum_{n=K+1}^{L} c_n \tilde{I}_{k+K-n}, \tag{21}$$

$$\tilde{y}^{(k)}_{k+n} = D(\tilde{y}^{(k-1)}_{k+n+1}) - c_{n+1}\tilde{I}_{k-1}, \quad n = 0, ..., K-1. \tag{22}$$

For sparse channels we propose the *thresholded complete feedback equalizer* TCFE

$$\tilde{y}^{(k)}_{k+K} = \tilde{y}_{k+K} = y_{k+K} - \sum_{n\in\Omega_c} c_n \tilde{I}_{k+K-n}, \tag{23}$$

$$\tilde{y}^{(k)}_{k+n} = D(\tilde{y}^{(k-1)}_{k+n+1}) - c_{n+1}\tilde{I}_{k-1}\delta_{\Omega_c}(n+1), \quad n = 0, ..., K-1, \tag{24}$$

$$\widehat{I}_k = \sum_{n=-K}^{0} f_n \tilde{y}^{(k)}_{k-n}, \tag{25}$$

$$\Omega_c = \{n : 1 \leq n \leq L, |c_n| > \eta_c\}. \tag{26}$$

where $\delta_{\Omega_c}(n) = 1$ if $n \in \Omega_c$ and 0 else. Note that all postcursor channel taps are thresholded prior to feedforward equalization so that the sparseness of $\{c_n\}$ can be fully exploited in the TCFE relative to the TPFE and TDFE. The TCFE requires $K + |\Omega_c| + 1$ H-CMULTs. If $\{c_n\}$ is sparse and $\{f_n\} * \{c_n\}$ is not sparse, we should have (for the same threshold) $|\Omega_c| \leq |\Omega_{p1}| + |\Omega_{p2}|$ in which case the TCFE requires fewer complex multiplies than the TPFE.

6 TPFE Initialization and Operation

The TPFE [3] is implemented as an LMS-type adaptive structure requiring a DFE update and a channel estimate update. The DFE update is governed by

$$\tilde{y}_{k+K}{=}y_{k+K} - \sum_{n\in\Omega p1} \hat{c}_n(k)\tilde{I}_{k+K-n} \tag{27}$$

$$f_n(k+1) = f_n(k) + 2u_1\tilde{y}^*_{k-n}\varepsilon_k, \quad n = -K, ..., 0 \tag{28}$$

$$b_{p,n}(k+1) = b_{p,n}(k) + 2u_1\tilde{I}^*_{k-n}\varepsilon_k, \quad n = 1, ..., K \tag{29}$$

where $\{\hat{c}_n\}$ signifies a channel estimate, and ε_k, and u_1 are the decision error, and adaptation stepsize respectively. It is only necessary to update $\{f_n\}$, $\{b_n\}$, and $\{\hat{c}_{\Omega p1,n}\}$ to obtain an adaptive TPFE. As the performance of the TPFE depends on the accuracy of

---

[3] TCFE adaptation follows quite naturally from the TPFE adaptation is is not discussed.

$\{\hat{c}_{\Omega,n}\}$, it is better to update $\{\hat{c}_n, n \in \Omega'\}$ to reduce the variance in the estimate, where $\Omega' \triangleq \{0, ..., L\}$, the indices of all channel taps. An initial channel estimate is computed off-line and used to initialize the TPFE structure. The idea is to select $\{\hat{c}_{\Omega,n}\}$ from $\{\hat{c}_n, n \in \Omega'\}$ to be used in the computation of the TPFE output decision. Given the availability of an initial channel estimate [4], the channel update proceeds as follows

$$\hat{y}_{k} = \sum_{n \in \Omega'} \hat{c}_n \tilde{I}_{k-n}, \tag{30}$$

$$\hat{c}_n(k+1) = \hat{c}_n(k) + 2u_2 \tilde{I}^*_{k-n} \gamma_k, \qquad n \in \Omega', \tag{31}$$

$$\hat{c}_{\Omega,n}(k+1) = \hat{c}_n(k+1), \qquad n \in \Omega. \tag{32}$$

where $\gamma_k = y_k - \hat{y}_k$, and u_2 is the adaptation stepsize.

The channel update lags the filter update by K samples. This is considered insignificant for slowly varying channels. Given that computationally expensive blind equalization must be addressed in any HDTV terrestrial broadcast receiver, it is reasonably assumed that a processor is available to perform (in software) the (sub-symbol rate) adaptive updates necessary on slowly varying channels. Our interest is in the reduction of complexity measured in CMULTs implemented in hardware for speed.

7 Static Simulation Results

The TPFE and TCFE were initially tested on seven test channels developed by the ATTC [5]. It was observed that precursor magnitude was a strong predictor of the equalizers' performance. Further tests were therefore run with variations of ATTC channel D (judged to be the worst) modified by changing its -20dB precursor to -10dB or -6dB. These three channels are referred to by their precursor magnitudes in this paper. Channel D is shown in Figure 3. Data was generated at a rate of 5.38Mhz from a 64-QAM[5] constellation and processed with shaping filters with 11.5% excess bandwidth. Adaptation misadjustment was avoided by utilizing theoretical filter solutions computed off-line using least squares.

The feedforward filter lengths for the structures were determined empirically based on the performance of the conventional DFE as the decision delay was varied. The feedback filter was chosen to cancel all postcursor ISI. This requirement results in a feedback filter length equal to the channel memory (126 taps). Performnce vs. decision delay (one less than feedforward filter length) is shown in Figure 4. The decision delays used in subsequent simulations are also indicated.

The output SNR of the TDFE, TPFE, and TCFE is plotted in Figure 5 as a function of the number of "feedback" taps, i.e., the combined channel and feedback filter taps ($|\Omega_d|$, $|\Omega_{p1} + \Omega_{p2}|$, $|\Omega_c|$, for the TDFE, TPFE and TCFE respectively). Simulations were conducted at 25dB receiver input SNR with 64-QAM for a target raw symbol-error rate of approximately 0.001 on the -10dB-precursor channel. Postcursor tap savings for the TPFE and TCFE relative to the TDFE were evaluated for 0.05dB and 0.5dB performance losses over the unthresholded DFE. The results are shown in Table 1. The TPFE yielded reductions in postcursor tap requirements of 22%-43% on the three channels.

---

[4] This is not very restrictive as the TPFE can be shown to converge to the correct solution regradless of an initial channel estimate if periodically available training symbols are utilized.

[5] results for 8-VSB are similar.

It is seen from Figure 5, and Table 1 that the TCFE performance is inferior to that of the TPFE on the -20dB-precursor channel. In this case, the advantages of full exploitation of the channel sparseness are offset by the use of generally non-sparse channel precursor taps in the TCFE structure. The TPFE and TCFE have comparable performance on these channels when a small threshold is used. The more consistent performance of the TPFE coupled with its simpler implementation may make it a more practical solution to sparse channel equalization.

8 Summary

Two reduced complexity DFE structures have been presented. Simulations demonstrate that on sparse channels, the TPFE and TCFE can yield large reductions in required H-CMULTs while maintaining performance comparable to the conventional DFE structure. Continuing research explores the use of pole zero modelling techniques applied to the TPFE and TCFE feedforward filters to further reduce complexity.

References

[1] J. G. Proakis, Digital Communications. New York: McGraw-Hill 1995

[2] B. Widrow and S.D. Stearns, Adaptive Signal Processing. Englewood Cliffs, NJ: Prentice-Hall 1985

[3] P. A. Voois, I. Lee, and J. Cioffi, "The effect of decision delay in finite length decision feedback equalization," IEEE Trans. Inform. Theory, vol. IT-42, pp. 618-621, Mar. 1996.

[4] D. D. Falconer, and F. R. Magee, "Adaptive channel memory truncation for maximum likelihood sequence estimation.," Bell Syst. Tech. J., vol. 52, pp. 1541-1562, Nov. 1973.

[5] W. F. Schreiber, "Advanced television systems for terrestrial broadcasting: some problems and proposed solutions," Proc. IEEE, vol. 83, pp. 958-981, June 1995.

[6] G. Young, "Reduced complexity decision feedback equalization for digital subscriber loops," IEEE J. Select. Areas Comm.,vol. 9, pp. 810-816, Aug. 1991.

[7] M. Ghavami, S. Nader-Esfahani, and W. Rupprecht, "Decision feedback equalization of HDSL using a generalized adaptive IIR delay filter," European Transactions on Telecommunications & Related Technologies, vol. 5, pp. 337-346, May-June 1994

[8] A. K. Aman, and S. A. Tretter, "A new type of decision feedback equalizer," Asilomar Conf. Circuits, Systems & Computers, vol. 1, pp. 265-269, 1991.

[9] S. Qureshi, and E. Newhall,"An adaptive receiver for data transmission over time-dispersive channels, " IEEE Trans. Inform. Theory,vol. IT-19, pp. 448-457, July 1973.

[10] N. Al-Dhahir, J. Cioffi, "MMSE decision-feedback equalizers: finite length results," IEEE Trans. Inform. Theory, vol. IT-41, pp. 961-975, July 1995

Precursor	Loss	TPFE	TCFE
-20dB	0.05dB	27%	13%
-10dB	0.05dB	29%	33%
-6dB	0.05dB	38%	41%
-20dB	0.5dB	22%	13%
-10dB	0.5dB	30%	31%
-6dB	0.5dB	42%	44%

Table 1: TPFE, TCFE vs TDFE tap reduction.

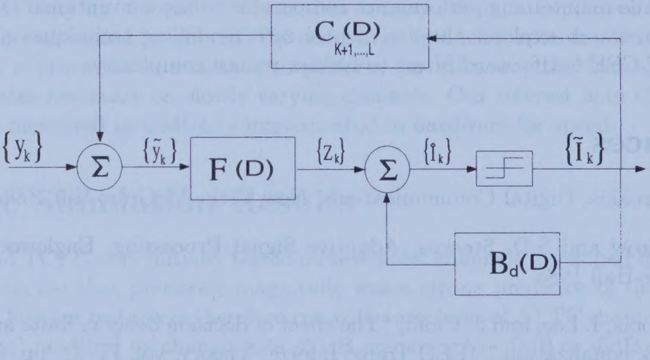

$C_{K+1,\ldots,L}(D)$
$\{y_k\}$
Σ
$\{\tilde{y}_k\}$
F (D)
$\{z_k\}$
Σ
$\{\hat{I}_k\}$
$\{\tilde{I}_k\}$
$B_d(D)$

Figure 1: Symbol spaced PFE structure.

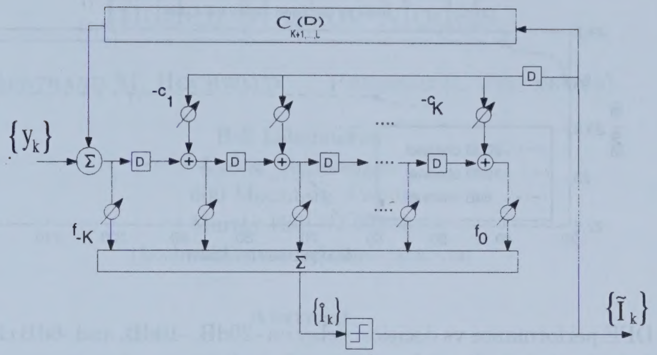

C(D) K+1,...,L
D
-c1
-cK
{y_k}
Σ
D
D
D
D
f-K
f0
Σ
{Î_k}
{Ĩ_k}

Figure 2: Symbol spaced CFE structure.

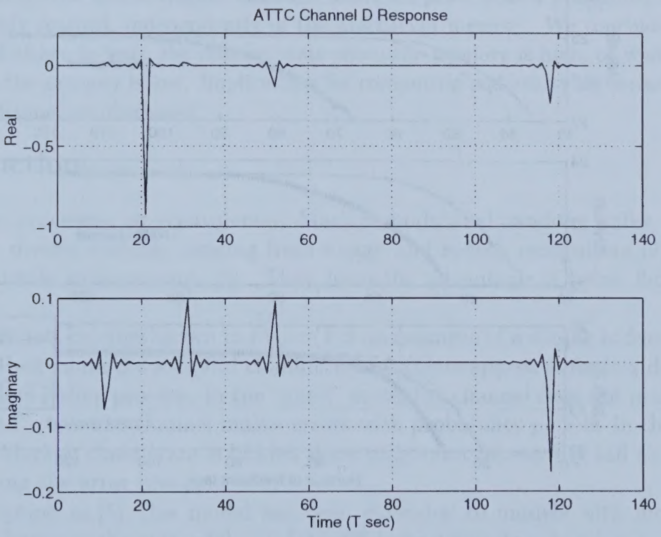


Figure 3: ATTC Channel D.

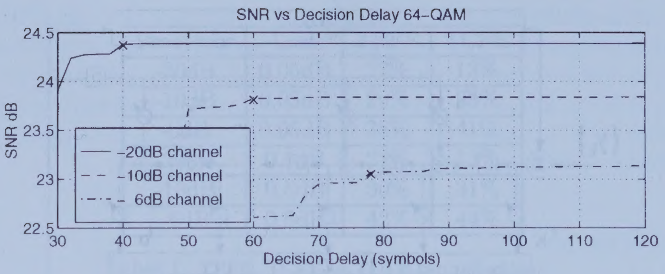


Figure 4: DFE performance vs decision delay on -20dB, -10dB, and -6dB channels. Delays used in simulations are indicated with x.

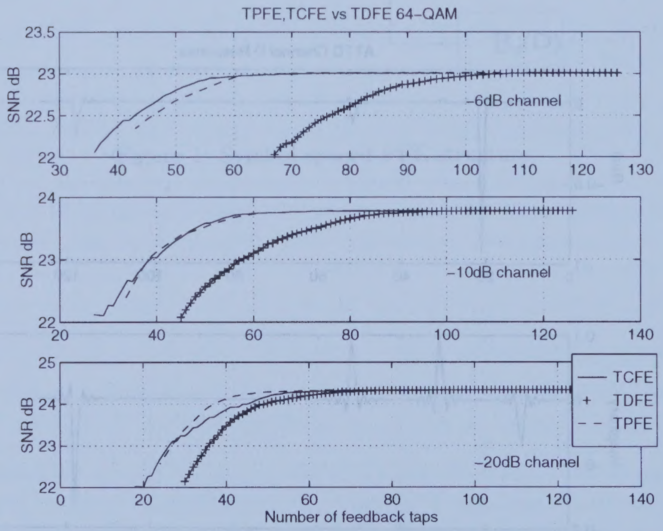


Figure 5: Performance vs number of equalizer non-zero postcursor taps for TDFE, TPFE, and TCFE: 64-QAM.

The Role of Memory in Entropy Approximation for Hidden Markov Models

BERTRAND M. HOCHWALD   PREDRAG R. JELENKOVIĆ

Bell Laboratories
Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974
{hochwald, predrag}@bell-labs.com

Abstract

Hidden Markov processes such as the Gilbert-Elliott channel have an infinite dependency structure. Therefore, entropy and channel capacity calculations require knowledge of the infinite past. In practice, such calculations are often approximated with a finite past. It is commonly assumed that the approximations require more of the past as the memory in the underlying Markov chain increases. We show that this is not necessarily true. We derive an exponentially decreasing upper bound on the accuracy of the finite-past approximation that is much tighter than existing upper bounds when the Markov chain mixes well. We also derive an exponentially decreasing upper bound that applies when the Markov chain does not mix at all. Our methods are demonstrated on the Gilbert-Elliott channel, where we prove that a prescribed finite-past accuracy is quickly reached, independently of the Markovian memory. We conclude that the past can be used either to learn the channel state when the memory is high, or wait until the states mix when the memory is low. Implications for computing and achieving capacity on the Gilbert-Elliott channel are discussed.

1. Introduction

Hidden Markov processes, or equivalently, Markov-modulated random walks, are used to model many diverse systems ranging from image and speech recognizers [9] to communication channels with memory [5]. They have the advantage of being flexible and simple.

The Gilbert-Elliott channel shown in Figure 1 is an example of a simple hidden Markov process that is used to model a digital channel whose errors appear in bursts due to, for example, a random fading process. In the "good" state, the channel does not make errors, while in the "bad" state the channel makes errors with probability $p_b > 0$. In this model, the underlying Markov chain state is hidden since we cannot necessarily tell the channel state by observing the error process.

Since its inception in [5], this model has been extended to models with more states and transitions between the states [3], [7], [11]. Of importance to achieving capacity on hidden Markov channels is the dependence of future errors on past ones, but even for the relatively simple two-state Gilbert-Elliott channel, this dependence is poorly understood. We show how, in entropy calculations, hidden Markov processes have dependence structures that are due to state "mixing" and "learning."

Loosely speaking, mixing is the act of switching states, which the underlying Markov chain does with frequency that depends on the off-diagonal elements of its transition

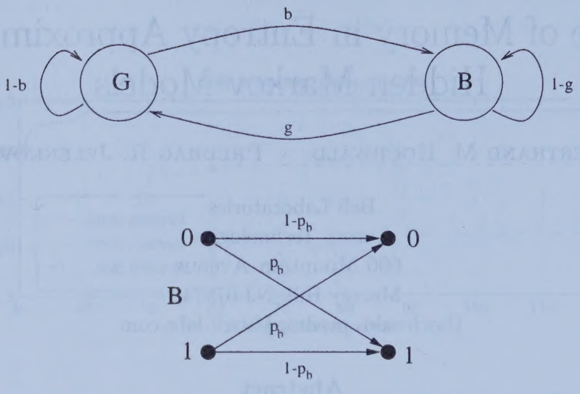


Figure 1: Hidden Markov model of Gilbert-Elliott channel

matrix. We prove that when the chain changes states often or mixes, the dependence structure is weak. On the other hand, when the chain mixes little, we are given a chance to learn the state from the observed process. We show that this makes the error dependence structure stronger, *but still limited.*

A Markov-modulated random walk is defined as a process $\{(S_n, Z_n),\ n = 0, 1, \dots\}$ such that

$$\mathbb{P}[S_n = j, Z_n = k | S_{n-1}, Z_{n-1}, \dots, S_0, Z_0] = \mathbb{P}[S_n = j, Z_n = k | S_{n-1}],$$

i.e., $\{(S_n, Z_n), n \geq 0\}$ is a Markov chain with past dependence on only the first coordinate. The transition probabilities for this chain are $p_{i,jk} = \mathbb{P}[S_n = j, Z_n = k | S_{n-1} = i] = q_{ij} a_{jk}$, where $q_{ij} \overset{\text{def}}{=} \mathbb{P}[S_n = j | S_{n-1} = i]$,and $a_{jk} \overset{\text{def}}{=} \mathbb{P}[Z_n = k | S_n = j, S_{n-1} = i]$. In this paper, we assume for simplicity that a_{jk} does not depend on i and that the number of states is finite. Abusing notation, we refer to Z_n as a Markov-modulated random walk if the accompanying underlying Markov chain S_n is self-understood.

For example, the error process Z_n of the Gilbert-Elliott channel is a Markov-modulated random walk. Here, Z_n is identically zero when S_n is the good state, and is one with probability p_b when S_n is the bad state. Letting the good state be denoted by a zero and the bad state with a one, we obtain $q_{00} = 1 - b$, $q_{01} = b$, $q_{10} = g$, $q_{11} = 1 - g$, $a_{00} = 1$, $a_{01} = 0$, $a_{10} = 1 - p_b$, and $a_{11} = p_b$. Thus, the modulating Markov chain has transition matrix

$$Q = \begin{pmatrix} 1-b & b \\ g & 1-g \end{pmatrix},$$

and has stationary probabilities $\pi_0 = g/(g+b)$ and $\pi_1 = b/(g+b)$. The transition matrix for the (S_n, Z_n) pair, for states (0,0), (1,0), and (1,1) is

$$P = \begin{pmatrix} 1-b & b(1-p_b) & bp_b \\ g & (1-g)(1-p_b) & (1-g)p_b \\ g & (1-g)(1-p_b) & (1-g)p_b \end{pmatrix}. \tag{1.1}$$

Throughout this paper, we assume that the transition probabilities b and g are parameterized by $\varepsilon \in [0,1]$ such that

$$b = b_0\varepsilon, \quad g = g_0\varepsilon, \qquad b_0,\, g_0 > 0. \tag{1.2}$$

The Markovian memory is defined as $\mu = 1 - b - g$, and we assume that b_0 and g_0 are chosen so that $\mu = 0$ when $\varepsilon = 1$. When $\mu = 0$ (or $\varepsilon = 1$) there is no Markovian memory and the channel errors $Z_0, Z_1, \ldots$, are independent and identically distributed. When $\mu = 1$ (or $\varepsilon = 0$) the chain is decomposable, but we assume that the stationary probabilities π_0 (good state) and π_1 (bad state) are defined. For simplicity, we do not consider negative memory $\mu < 0$ or $\varepsilon > 1$, but our arguments can easily be generalized to include this case.

Let $\{S'_n, n \geq 0\}$ be a stationary Markov chain with transition probability $P = \{p_{ij}\}$ and finite state space $E = \{1, \ldots, A\}$. If f is a function that maps $E \to \{1, \ldots D\}$, then the process $Z = \{Z_n \overset{\text{def}}{=} f(S'_n), n \geq 0\}$ is called a *function of the Markov chain S'*.

Functions of a Markov chain and Markov-modulated random walks are equivalent. That a Markov-modulated random walk is a function of a Markov chain follows immediately from the definition, since $Z_n = g(S_n, Z_n)$, where $g(\cdot,\cdot)$ is the function that takes the second argument. On the other hand, suppose that Z_n is a function of a Markov chain; that is, $Z_n = f(S'_n)$ for some function $f(\cdot)$ of the Markov chain S'_n. Then, it is plain that the pair (S'_n, Z_n) form a Markov modulated random walk with $a_{jk} = 1$ when $k = f(j)$, and $a_{jk} = 0$ otherwise.

Let Z be a function of a stationary Markov chain. The entropy of Z is defined as $H(Z) \overset{\text{def}}{=} \lim_{n\to\infty}(1/n)H(Z^{(n)})$ (whenever this limit exists), where $Z^{(n)} \overset{\text{def}}{=} \{Z_1, \ldots, Z_n\}$. Among other roles, $H(Z)$ is fundamental towards determining the capacity of channels with hidden Markov structure [4].

This limit is generally difficult to compute and is often approximated by the following upper and lower bounds,

$$H(Z_n|Z^{(n-1)}, S_0) \leq H(Z) \leq H(Z_n|Z^{(n-1)}), \tag{1.3}$$

where

$$\lim_{n\to\infty} H(Z_n|Z^{(n-1)}, S_0) = H(Z) = \lim_{n\to\infty} H(Z_n|Z^{(n-1)}). \tag{1.4}$$

The proof of these results can be found in [1] or [2, pg. 63]. For numerical computation, one can use either $H(Z_n|Z^{(n-1)}, S_0)$ or $H(Z_n|Z^{(n-1)})$ for some finite n to approximate $H(Z)$. It is of practical importance to know the error of either approximation, but rather than work with these quantities separately, we combine them and define the conditional mutual information $I(Z_n; S_0|Z^{(n-1)}) \overset{\text{def}}{=} H(Z_n|Z^{(n-1)}) - H(Z_n|Z^{(n-1)}, S_0) \geq 0$. Equation (1.4) says that $I(Z_n; S_0|Z^{(n-1)}) \to 0$ as $n \to \infty$; that is, Z_n and S_0 become independent as $n \to \infty$, conditionally on the $(n-1)$th order history. This paper is concerned primarily with the rate with which this happens.

It is commonly believed that $I(Z_n; S_0|Z^{(n-1)})$ converges to zero with a rate that goes to zero as q_{ij} $(i \neq j)$ tends to zero (for example, as $\varepsilon \to 0$ in the Gilbert-Elliott channel and the underlying Markov chain becomes decomposable). Theorem 1, given in the next section and first derived in [1], strengthens this belief by finding an upper bound on $I(Z_n; S_0|Z^{(n-1)})$ for functions of a Markov chain that converges exponentially as ζ^n for some ζ that, in general, approaches one as q_{ij} $(i \neq j)$ tends to zero. Define $-\log\zeta$ to be the *rate* of the bound and define $\lim_{n\to\infty} -(1/n)\log I(Z_n; S_0|Z^{(n-1)})$ to be the true *convergence rate* (assuming that this exists). Then Theorem 1 gives a convergence rate lower bound that approaches zero as $q_{ij} \to 0$ $(i \neq j)$.

However, in this paper we argue that Theorem 1 gives a very loose bound and there are, instead, two exponential rates of convergence, in general. One is due to mixing and dominates when the connection probabilities are nonzero, and the other is due to learning, when the connection probabilities are zero. *Neither rate is zero.* We demonstrate our arguments explicitly on the Gilbert-Elliott channel.

2. Bibliographical Note on the Convergence Rate

The best known bound on the rate of convergence of $I(Z_n; S_0|Z^{(n-1)})$ is due to Birch [1].

Theorem 1

$$I(Z_n; S_0|Z^{(n-1)}) \le B\zeta^{n-1}, \tag{2.1}$$

where

$$\zeta = 1 - \frac{N_1}{N_D^2} \min_{i,j,k,m,n} \left(\frac{p_{ik}p_{kn}}{p_{ij}p_{jm}}\right)^2, \qquad B = (N_D \log e)/(N_1 \min_{i,j} p_{ij}),$$

and N_1 and N_D are the minimum and maximum number of states collapsed by f.

While exponential rates of convergence are often desirable, the rate $-\log\zeta$ in Theorem 1 is, in general, very weak and suggests that a very large history is needed. We demonstrate this with the Gilbert-Elliott channel.

As described in Section 1, this channel can be described as the function that takes the second coordinate of the Markov chain with transition matrix given by (1.1). To apply the theorem, we note that $N_1 = 1$ and $N_D = 2$. Let the transition probabilities b and g of the Markov chain S_n be parameterized as in (1.2) and consider $\varepsilon \to 0$. Then $\zeta = 1 - O(\varepsilon^4)$ and $B = \Omega(1/\varepsilon)$ and

$$I(Z_n; S_0|Z^{(n-1)}) \le B[1 - O(\varepsilon^4)]^n$$

where B goes to infinity as $\varepsilon \to 0$. Hence, the rate is $-\log\zeta = -\log(1 - O(\varepsilon^4)) = O(\varepsilon^4)$, which rapidly goes to zero as the Markovian memory in S_n rises. With its dependency on ε^4, this lower bound on the convergence rate turns out to be extremely loose. In Section 4, we show that the true convergence rate for small ε is, in fact, *uniformly bounded* from below.

At the other extreme, as $\varepsilon \to 1$ the Markov chain states $S_0, S_1, \ldots,$ and the channel errors $Z_0, Z_1, \ldots,$ become, within themselves, independent and identically distributed processes. Therefore, $I(Z_n; S_0|Z^{(n-1)}) \to 0$ as $\varepsilon \to 1$ for every n, and the true convergence rate should go to infinity. However, ζ in Theorem 1 does not go to zero ($-\log\zeta$ does not go to infinity) as $\varepsilon \to 1$, and thus the bound (2.1) again becomes arbitrarily loose.

In the next section we derive a new much tighter bound that is as widely applicable as Theorem 1. We show that the rate of our new bound approaches infinity as $\varepsilon \to 1$ in the Gilbert-Elliott channel. The bound is derived with the help of a contraction coefficient first considered by G. Birkhoff. Our novel application of this contraction coefficient to conditional entropy is, we believe, of independent interest.

3. Hilbert's Metric and Birkhoff's Contraction Coefficient

In order to state and prove our results we need to introduce the notion of Hilbert's pseudo-metric [10, pg. 80] d which is defined for any two vectors $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n)$ with positive elements as

$$d(x, y) = \max_{i,j} \log \left(\frac{x_i y_j}{x_j y_i} \right).$$

This pseudo-metric has a property that $d(x, y) = 0$ if and only if $x = \lambda y$ for some scalar λ. The metric turns out to be ideally suited for analytical manipulations involving conditional entropy.

For any matrix T with nonnegative elements we define Birkhoff's ergodic coefficient as

$$\tau(T) = \sup_{x,y>0, x \neq y} \frac{d(xT, yT)}{d(x, y)}.$$

An explicit formula for $\tau(T)$ is given as

$$\tau(T) = \frac{1 - \sqrt{\phi(T)}}{1 + \sqrt{\phi(T)}}, \qquad \phi(T) = \min_{i,j,k,l} \frac{t_{ik} t_{jl}}{t_{jk} t_{il}}.$$

This is surprisingly tedious to prove [10]. It is clear that $\tau(T) = 0$ if and only if T is of rank one.

3.1 Convergence Rate

To reduce the notational complexity, we consider the special case where the function f of the Markov chain partitions the state space E into sets of equal size; i.e., we assume that the sets $E_i = f^{-1}(i) \overset{\text{def}}{=} \{j : f(j) = i\}$ all have K elements, where $KD = A$. Without loss of generality suppose that the state space is labeled such that $f(j) = i$, for all $(i-1)K < j \leq iK$, i.e., $E_i = \{j : (i-1)K < j \leq iK\}$. Then, we can partition the transition matrix P into blocks P_{ij}

$$P_{ij} = \{p_{kl} \, : \, k \in E_i, l \in E_j\}, \quad i, j = 1, \dots, D.$$

Let π be the stationary probability vector for the Markov chain P, and let $\pi^i = (\pi_{(i-1)K+1}, \dots, \pi_{iK})$, $i = 1, \dots, D$ be the subvectors corresponding to the subsets E_i. Similarly, let us break the jth row of the matrix P into D blocks $p^i_{j\cdot} = (p_{j,(i-1)K+1}, \dots, p_{j,iK})$, $i = 1, \dots, D$, each of length K.

Theorem 2

$$I(Z_n; S_0 | Z^{(n-1)}) \leq C \rho^{n-2}, \qquad n \geq 2, \tag{3.1}$$

where

$$\rho = \max_{i,j} \tau(P_{ij}), \qquad C = \max_{1 \leq i \leq A, \, 1 \leq j \leq D} d(p^i_{j\cdot}, \pi^j).$$

Proof: Given in [6]. ◇

3.2 Application to Gilbert-Elliott Channel

Let us now compare bounds (2.1) and (3.1). As $\varepsilon \to 1$ ($\mu \to 0$), we already know from Section 2 that the true convergence rate goes to infinity whereas the rate given by Theorem 1 remains bounded. Applying Theorem 2 to the transition matrix (1.1) yields

$$\phi = \frac{1}{1+\frac{\mu}{bg}}, \qquad \rho = \frac{\sqrt{1+\frac{\mu}{bg}}-1}{\sqrt{1+\frac{\mu}{bg}}+1},$$

and it is seen that $-\log\rho \to \infty$ as $\mu \to 0$. Thus, Theorem 2 gives a much tighter bound than Theorem 1.

On the other hand, as $\varepsilon \to 0$ we have $\phi = O(\varepsilon^2)$ and $\rho = 1 - O(\varepsilon)$. Therefore, the rate $-\log\rho = O(\varepsilon)$ depends linearly on the Markovian memory and the bound given in Theorem 2 is again much tighter than the bound given in Theorem 1. The linear dependence on ε also seems intuitively reasonable since the Markovian memory is $\mu = 1 - O(\varepsilon)$. However, as we now show, there exists a uniform bound on the convergence rate that is independent of ε and therefore, the bound in Theorem 2 also inevitably becomes loose as $\varepsilon \to 0$.

4. Universal Bounds

While Theorem 2 provides a tight bound when q_{ij} ($i \neq j$) are not too small (ε not too small in Gilbert-Elliott channel), we now show that the rate of convergence *is nonzero* even when $q_{ij} = 0$ ($i \neq j$). Roughly speaking, Theorem 2 captures the rate due to state mixing, but when $q_{ij} = 0$ ($i \neq j$), the convergence is instead due to state learning. Hence, Theorem 2, and especially Theorem 1, inevitably provide loose bounds for small q_{ij} ($i \neq j$).

We begin by deriving universal bounds for the Gilbert-Elliott channel before proceeding to a learning bound for general functions of Markov chains. In the language of this channel, we wish to bound the speed with which $I(Z_n; S_0|Z^{(n-1)})$ converges, uniformly in ε. That is, we want to show that a fixed amount of history is needed, no matter how large the Markovian memory (or how small ε) is.

4.1 Universal Mixing Bound for Gilbert-Elliott Channel

In this section we assume that $\varepsilon > 0$, and that the Markov chain therefore mixes. Define

$$x_{ij}^n = \mathbb{P}[S_0 = i, Z_1 = 0, \ldots, Z_{n-1} = 0, Z_n = j] \quad i, j \in \{0, 1\}.$$

It is easy to see that x_{ij}^n has the form

$$x_{ij}^n = c_{ij}\lambda_1^{n-1} + d_{ij}\lambda_2^{n-1}, \tag{4.1}$$

for some c_{ij} and d_{ij}, where $0 < \lambda_1 < \lambda_2 < 1$ are the eigenvalues of the upper left-hand 2×2 submatrix of P in (1.1)

$$P_{00} = \begin{bmatrix} 1-b & bq_b \\ g & (1-g)q_b \end{bmatrix}, \qquad q_b \overset{\text{def}}{=} 1 - p_b.$$

By carefully analyzing x_{ij}^n in [6], we prove the following theorem.

Theorem 3 *Let $\varepsilon > 0$ and $\lambda_1 < \lambda_2$ be the eigenvalues of P_{00}. Then the conditional mutual information for the Gilbert-Elliott channel obeys*

$$I(Z_n; S_0|Z^{(n-1)}) \sim C\left(\frac{(\lambda_1)^2}{\lambda_2}\right)^{n-1}$$

as $n \to \infty$. The constant C is given explicitly in [6].

Proof: Given in [6]. ◇

It is straightforward to show that $(\lambda_1)^2/\lambda_2 \nearrow (1-p_b)^2$ as $\varepsilon \searrow 0$. Hence, when $\varepsilon > 0$, or equivalently, when the underlying Markov chain mixes, the convergence rate of $I(Z_n; S_0|Z^{(n-1)})$ is uniformly bounded by $-2\log(1-p_b)$, *independently of the Markovian memory.*

4.2 Universal Learning Bound for Gilbert-Elliott Channel

We have shown in the preceding section that the the convergence rate of $I(Z_n; S_0|Z^{(n-1)})$ due to mixing is given by a quantity that approaches $-2\log(1-p_b)$ as $\varepsilon \to 0$. From this, we might be tempted to infer that the convergence rate is exactly $-2\log(1-p_b)$ when $\varepsilon = 0$. However, the analysis in the preceding section is invalid when $\varepsilon = 0$. We now show that the convergence rate is, in fact, $-\log(1-p_b)$ rather than $-2\log(1-p_b)$, and that this rate is due to learning rather than mixing. Clearly $-\log(1-p_b) < -2\log(1-p_b)$ and therefore $-\log(1-p_b)$ is a lower bound on the convergence rate, universally valid for all $\varepsilon \geq 0$.

Before we state the principle result, we generalize the Gilbert-Elliott channel to allow channel errors when in the good state. We assume that, when in a good state, the channel is binary symmetric with error probability $p_g \geq 0$.

Theorem 4 *Define*

$$D(p_1, p_2) \overset{\text{def}}{=} p_1 \log\frac{p_1}{p_2} + (1-p_1)\log\frac{1-p_1}{1-p_2},$$

and let $\varepsilon = 0$, and p_g, $p_b > 0$. Then

$$\lim_{n\to\infty} \frac{-\log I(Z_n; S_0|Z^{(n-1)})}{n} = D(p^*, p_b) = D(p^*, p_g), \tag{4.2}$$

where

$$p^* \overset{\text{def}}{=} \frac{\log\frac{1-p_g}{1-p_b}}{\log\frac{p_b}{p_g} + \log\frac{1-p_g}{1-p_b}}, \qquad p_g < p^* < p_b.$$

Proof: Given in [6]. ◇

Remark: When $p_g = 0$ (error-free good state) the expression for $I(Z_n; S_0|Z^{(n-1)})$ is especially simple and we may obtain its exact asymptotic form. In this case

$$I(Z_n; S_0|Z^{(n-1)}) \sim \pi_0 p_b \log\left(\frac{1}{1-p_b}\right) n(1-p_b)^{n-1}.$$

Hence the convergence rate of $I(Z_n; S_0|Z^{(n-1)})$ for $\varepsilon = p_g = 0$ is $-\log(1-p_b)$. This concurs with Theorem 4 since $D(p^*, p_b) \to -\log(1-p_b)$ as $p_g \to 0$.

We now connect the conclusion of Theorem 4 with state-learning. Suppose that $\varepsilon = 0$ and we would like to estimate $S_n = S_0$ from $Z^{(n-1)}$. We could count the number of ones in the sequence $Z_1, \ldots, Z_{n-1}$, divide by n, and compare this value to a threshold p. Clearly, any value for p such that $p_g < p < p_b$ will yield an estimate that converges in probability to the correct state as $n \to \infty$. It turns out that the convergence speed of the estimate is optimized when $p = p^*$.

To show this, we define n_z as the number of ones in $Z_1, \ldots, Z_n$, and define the estimate $\hat{S}_0$ as

$$\hat{S}_0 = \begin{cases} 0 \text{ (good state)} & \text{if } n_z/n \le p \\ 1 \text{ (bad state)} & \text{if } n_z/n > p \end{cases}$$

Then the probability that the estimate fails is

$$\mathbb{P}[\hat{S}_0 \neq S_0] = \pi_0 \mathbb{P}[n_z > np | S_0 = 0] + \pi_1 \mathbb{P}[n_z \le np | S_0 = 1].$$

Given $S_0 = 0$, the random variable n_z has a binomial distribution with parameters n and p_g; and given $S_0 = 1$, n_z is binomial with parameters n and p_b. A standard application of Cramér's theorem in large deviation theory (see, for example, [12]) yields

$$\lim_{n\to\infty} \frac{-\log \mathbb{P}[\hat{S}_0 \neq S_0]}{n} = \min(D(p, p_g), D(p, p_b)).$$

As functions of p, $D(p, p_g)$ is monotonically increasing for $p \ge p_g$, and $D(p, p_b)$ is monotonically decreasing for $p \le p_b$. The convergence rate of the estimate to the correct answer is therefore maximized by choosing τ so that $D(p, p_g) = D(p, p_b)$. Thus, $p = p^*$ and the learning rate is $D(p^*, p_g) = D(p^*, p_b)$.

4.3 General Learning Bound

The previous section shows that it is possible to get an exponential rate of convergence even when the underlying Markov chain does not mix at all. The convergence is due to the effects of state learning that are overlooked by Theorems 1 and 2. We now briefly generalize the concept of learning to arbitrary Markov-modulated random walks.

We assume that Z is a Markov-modulated random walk such that $q_{ij} = 0$ for all $i \neq j$, and $(a_{j1}, a_{j2}, \ldots) \neq (a_{j'1}, a_{j'2}, \ldots)$ for all $j \neq j'$, where q_{ij} and $a_{j1}, \ldots$ are defined in the introduction. That is, the underlying Markov chain does not mix, and the discrete state $S_n = S_0$ uniquely determines the discrete distribution of Z_n. For example, these requirements become $\varepsilon = 0$ and $p_g \neq p_b$ when applied to the Gilbert-Elliott channel. We obtain the following theorem.

Theorem 5 *Let Z be a Markov-modulated random walk such that $q_{ij} = 0$ for all $i \neq j$, and $(a_{j1}, a_{j2}, \ldots) \neq (a_{j'1}, a_{j'2}, \ldots)$ for all $j \neq j'$. Let $\hat{S}_0 = \hat{S}_0(Z^{(n-1)})$ be an estimate of S_0 that converges in probability exponentially quickly to S_0 with some rate $r > 0$. Then*

$$\limsup_{n\to\infty} \frac{\log I(Z_n; S_0|Z^{(n-1)})}{n} \le -r.$$

Proof: Given in [6]. ◇

Remark: The estimate $\hat{S}_0$ that gives the largest r (fastest learning) gives the tightest bound.

5. Implications for Gilbert-Elliott Channel

The capacity of the Gilbert-Elliott channel is often computed by using (1.3) to approximate $H(Z)$ to some prescribed accuracy. Suppose that we wish to estimate how large n needs to be so that $H(Z_n|Z^{(n-1)})$ is within 0.01 of $H(Z)$. A large estimate for n would be discouraging since the computational complexity of either $H(Z_n|Z^{(n-1)})$ or $H(Z_n|Z^{(n-1)}, S_0)$ grows as 2^n. Let the Gilbert-Elliott parameters be, for example, $b = 0.1$, $g = 0.5$, $p_g = 0.01$ and $p_b = 0.5$. Then $B \approx 289$ and $\zeta \approx 1 - 10^{-8}$ in Theorem 1, and thus the estimate according to this theorem is $n \approx 10^9$ for $I(Z_n; S_0|Z^{(n-1)}) \approx 0.01$. This means that, according to Theorem 1, approximately 2^{10^9} floating point operations are needed, seemingly putting an accurate estimate of $H(Z)$ well beyond our reach. Furthermore, observe that this estimate of n goes to infinity as $p_g \to 0$ because $\zeta \to 1$.

However, our bound in Theorem 2 yields $I(Z_n; S_0|Z^{(n-1)}) \leq 5.49 \cdot 0.5^{n-2}$ independently of p_g. Therefore, $n = 9$ suffices. This much more reasonable value of n is well within our computationally ability.

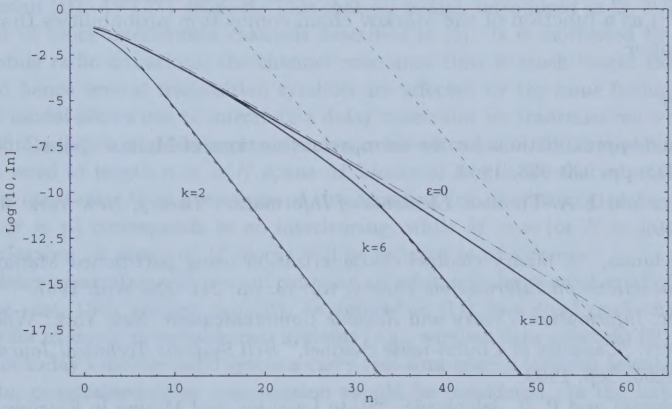


Figure 2: Graph of $\log_{10} I(Z_n; S_0|Z^{(n-1)})$ versus n for $p_g = 0$. Solid lines represent $I(Z_n; S_0|Z^{(n-1)})$ for $p_b = 0.5$, $g = 5 * 10^{-k}$, $b = 10^{-k}$ for $k = 2$, 6, and 10. Long-dashed line is $\varepsilon = 0$ asymptote (learning rate $-\log(1 - p_b)$ given by Theorem 4), while short-dashed lines are given by Theorem 3.

Figure 2 demonstrates the mixing and learning rates for $p_g = 0$. For small n, these curves have decay rates that are dominated by learning and are well approximated by the $\varepsilon = 0$ learning asymptote (given by long-dashed line and Theorem 4), while for larger n, $I(Z_n; S_0|Z^{(n-1)})$ decays at the mixing rate given in Theorem 3. Observe the rather abrupt change in rate (knees in the curves) as n increases. From the figure, we see that the mutual information is always bounded by the $\varepsilon = 0$ learning asymptote. Therefore, for $n = 20$, $I(Z_n; S_0|Z^{(n-1)})$ is approximately 10^{-5}, independently of ε.

To achieve capacity on the Gilbert-Elliott channel, Mushkin and Bar-David in [8] propose a scheme involving a $J \times N$ dimensional interleaver and similar deinterleaver.

They prove that the capacity achievable on the jth output of the J-output channel is given by $C^{(j)} = 1 - H(Z_j|Z^{(j-1)}), j = 1, \dots, J$. The capacity of the Gilbert-Elliott channel itself is given by $C^* \overset{\text{def}}{=} \lim_{j\to\infty} C^{(j)}$. Our theoretical results show that, regardless of the channel memory, J does not need to be large for $C^{(J)}$ to be close to C^*.

6. Conclusion

We have argued that, in general, $I(Z_n; S_0|Z^{(n-1)})$ decreases exponentially with n and therefore $H(Z_n|Z^{(n-1)})$ and $H(Z_n|Z^{(n-1)}, S_0)$ both approach the limiting entropy $H(Z)$ exponentially quickly. We identified mixing and learning rates for functions of a Markov chain and showed that the worst-case convergence for the Gilbert-Elliott channel was given by the explicitly identified learning rate. Learning rates for other functions of Markov chains can also be computed using the techniques we have outlined. These rates are generally nonzero as long as the observed process has a distribution that is uniquely determined by the underlying state.

Based on the evidence we have given, we conjecture that, in general, the worst case convergence of $I(Z_n; S_0|Z^{(n-1)})$ is obtained when $q_{ij} = 0$ $(i \neq j)$. Hence the worst-case convergence is still exponential, but at the learning rate. The amount of past needed to obtain a prescribed accuracy in (1.3) can thus be chosen independently of q_{ij}. Proving (or disproving) this conjecture will probably involve going deeper into the behavior of $I(Z_n; S_0|Z^{(n-1)})$ as a function of the Markov chain connection probabilities than we have done in this paper.

References

[1] J. J. Birch, "Approximations for the entropy for functions of Markov chains," *Ann. Math. Stat.*, vol. 33, pp. 930–938, 1962.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: John Wiley & Sons, 1991.

[3] B. D. Fritchman, "A binary channel characterization using partitioned Markov chains," *IEEE Transactions on Information Theory*, vol. 13, pp. 221–226, Apr. 1967.

[4] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[5] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Systems Technical Journal*, vol. 39, pp. 1253–1265, Sep. 1960.

[6] B. M. Hochwald and P. R. Jelenković, "State Learning and Mixing in Entropy of Hidden Markov Processes and the Gilbert-Elliot Channel," *Bell Labs Technical Memorandum*, BL0112170-970904-15TM, 1997.

[7] L. N. Kanal and A. R. K. Sastry, "Models for channels with memory and their applications to error control," *Proceedings of the IEEE*, vol. 66, July 1978.

[8] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliot channels," *IEEE Transactions on Information Theory*, vol. 35, pp. 1277–1290, Nov. 1989.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[10] E. Seneta, *Non-negative Matrices and Markov Chains*. Springer-Verlag, 1981.

[11] W. Turin and M. M. Sondhi, "Modeling error sources in digital channels," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 340–347, Apr. 1993.

[12] A. Weiss and A. Shwartz, *Large Deviations for Performance Analysis: Queues, Communications, and Computing*, New York: Chapman & Hall, 1995.

Power control for the fading channel:
An information-theoretical approach

Giuseppe Caire Giorgio Taricco Ezio Biglieri*

Abstract

We evaluate the information outage rate of the block-fading Gaussian channel with constant-rate transmission and a delay constraint, under the assumption that both transmitter and receiver have perfect channel-state information. The power allocation scheme that minimizes the outage rate for a given average transmitted power is derived. An iterative algorithm for optimal power allocation, which needs no *a priori* knowledge of the fading statistics, is also exhibited, Applications to the Rayleigh fading channel are developed. In particular, a closed-form expression for the delay-limited capacity of the 2-block Rayleigh fading channel is derived.

I INTRODUCTION

The main goal of this paper is the derivation of optimal power-control policies for the block-fading Gaussian (BF-AWGN) channel. This channel model, introduced in [2, 3], belongs to the general class of block-interference channels described in [5]. It is motivated by the fact that, in many mobile radio situations, the channel coherence time is much longer than one symbol interval, and hence several transmitted symbols are affected by the same fading value. Use of this channel model allows one to introduce a delay constraint for transmission, which is realistic whenever infinite-depth interleaving is not a reasonable assumption. With this model, we assume that a code word of length $n = MN$ spans M blocks of length N; the value of the fading in each block is constant. M is a measure of the decoding (or interleaving) *delay* of the system: $M = 1$ (or $N = n$) corresponds to no interleaving, while $M = n$ (or $N = 1$) corresponds to perfect interleaving. A group of M blocks will be referred to as a *frame*.

With no delay constraint, a code word can span an arbitrarily large number M of fading blocks. If this is the case, then average capacity, as derived in [11], is a good performance indicator. This applies for example to variable-rate systems (e.g., wireless data networks [9]). On the other hand, most of today's mobile radio systems carry real-time speech (cellular telephony), for which constant-rate, constrained-delay transmission should be considered. In the latter case outage rate, rather than average capacity, is the appropriate performance limit indicator [2, 4].

Here we solve the problem of minimizing the outage rate of the M-block BF-AWGN under the assumption that both transmitter and receiver have perfect CSI. We show that minimum outage rate can be achieved by transmitting a fixed code book, randomly generated with i.i.d. Gaussian components, and by suitably allocating the transmitted power to the M blocks. The optimal power-allocation policy is derived under a constraint on the average transmitted power. It is found to depend on a threshold, which in turn depends on the fading statistics. Since these may be unknown a priori, we also exhibit a simple algorithm that estimates this threshold without prior knowledge of the fading statistics, and can be made adaptive so as to track slowly-varying fading statistics.

A quantity related to the outage rate is the *delay-limited capacity*. This was defined in [12] in the case of a fading multiple access channel (MAC). In the multiuser case with $M = 1$, the delay-limited capacity region was solved by using the polymatroid structure of the capacity region of the Gaussian MAC. When $M > 1$, the latter structure cannot be exploited even in

---

*Dipartimento di Elettronica • Politecnico • Corso Duca degli Abruzzi 24 • I-10129 Torino • fax: +39 11 5644099 • e-mail: `<name>@polito.it` • This work was supported by CNR under "Progetto Finalizzato Trasporti."

the simple single-user case. Here we re-state the definition of delay-limited capacity in terms of the M-block BF-AWGN channel, and show that solving the minimum outage rate problem is tantamount to deriving the delay-limited capacity. In particular, an explicit formula is given for the delay-limited capacity of the independent Rayleigh BF-AWGN channel with $M = 2$ blocks. This shows that a small delay ($M = 2$) can buy a considerable increase in transmission reliability (the delay-limited capacity is zero for $M = 1$). Moreover, the delay-limited capacity of this channel turns out to be only 5 dB away from the capacity of the AWGN channel, and 2.5 dB away from the average capacity of the Rayleigh fading channel, which correspond to a very large delay ($M \to \infty$).

The paper is organized as follows. In Section II we define the BF-AWGN, we recall known results about its average capacity, and we formally introduce the outage rate and the delay-limited capacity. In Section III we solve the minimum outage rate problem, while in Section IV some numerical examples for the Rayleigh fading channel are presented and discussed. Finally, in Section V, we point out our conclusions.

Since the proofs of our results are rather long and technical, they are omitted here. However, they can be found in a full paper now in preparation [15] and available from the authors.

II SYSTEM MODEL, CAPACITY, AND OUTAGE RATE

The discrete-time real block-fading additive white Gaussian noise (BF-AWGN) channel is defined by

$$y[k] = \sqrt{\alpha[k]}x[k] + n[k] \tag{1}$$

where $n[k]$ (the noise) is a sequence of i.i.d. $\sim \mathcal{N}(0, 1)$ random variables, $x[k]$ (the information symbols) is a sequence of real random variables, and $\alpha[k]$ (the channel power gain) is a sequence of non-negative random variables independent of $x[k]$ and $n[k]$, piecewise constant over N adjacent symbols. Without loss of generality, we let

$$\alpha[mN + n] = \alpha_m \quad \text{for} \quad n = 0, \ldots, N - 1 \text{ and } m \in \mathbb{Z} \tag{2}$$

The sequence $\{\alpha_m\}$ is assumed to be ergodic (although not necessary i.i.d.) [2, 3]. The transmission of a code word of length MN is time-aligned with the fading, so that a code word spans exactly M channel blocks of length N for which the fading is constant [5].

As discussed in Section I, the delay constraints are reflected by the choice of a finite value for M.

Average capacity. When $N < \infty$ and M is unconstrained, the average capacity of the BF-AWGN channel under the input power constraint $\mathrm{E}[x^2[k]] \leq \mathcal{P}$ is given by (see [2, 3, 5, 11])

$$C_{\text{constr}}(\mathcal{P}) = \mathrm{E}\left[\frac{1}{2}\log(1 + \alpha\mathcal{P})\right] \quad \text{nat/symbol} \tag{3}$$

when only the receiver has perfect CSI, and by

$$C_{\text{pow}}(\mathcal{P}) = \max_{\gamma} \mathrm{E}\left[\frac{1}{2}\log(1 + \alpha\gamma)\right] \quad \text{nat/symbol} \tag{4}$$

when both transmitter and receiver have perfect CSI. In the expressions above, E denotes expectation taken wrt the random variable α. The optimal transmitted-power allocation in the first case is constant [3]. In the second case, the optimal function $\gamma(\alpha)$ is derived in [11].

Outage rate and delay-limited capacity. We define the M-block "instantaneous" mutual information $I_M(\alpha, \gamma)$ as the mutual information per symbol resulting from the transmission over M blocks, conditionally on the channel gains $\boldsymbol{\alpha} = \{\alpha_m\}_{m=0}^{M-1}$ and on the piecewise-constant power allocation $\boldsymbol{\gamma} = \{\gamma_m\}_{m=0}^{M-1}$.[1]

Assume the CSI $\boldsymbol{\alpha}$ to be known at both transmitter and receiver. With given $\boldsymbol{\alpha}$ the channel becomes Gaussian, and hence the maximum information transfer is achieved by assuming independent (conditionally on $\boldsymbol{\gamma}$) Gaussian inputs $x[mM+n] \sim \mathcal{N}(0, \gamma_m)$ (see [2] for a formal proof). Thus, we can write

$$I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{M} \sum_{m=0}^{M-1} \frac{1}{2} \log(1 + \alpha_m \gamma_m) \tag{5}$$

For $M < \infty$, $I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ is a random variable. Consider the case of transmission at constant rate R nat/symbol, and assume a very large N. Although long code words can be transmitted (recall that the code word length is $n = MN$), the Shannon capacity of the M-block BF-AWGN channel might be zero, since $I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ can be below any $R > 0$ with a probability that does not vanish as $N \to \infty$ [2, 3].

In this context, a sensible parameter expressing the quality of the channel is the *information outage rate*: this is the probability that $I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ falls below the rate R [2], that is,

$$P_{\text{out}}(R) = P(I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma}) < R) \tag{6}$$

The operational significance of $P_{\text{out}}(R)$ is given by the following result [14, 15]:

Proposition 1. $P_{\text{out}}(R)$ is equal to the word-error probability averaged over the random codes of rate R and length $n = MN$ with i.i.d. components $\sim \mathcal{N}(0,1)$, and over the joint realizations of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ for M fixed and $N \to \infty$.

Comment. Proposition 1 states that if we take good codes (i.e., codes for which the word error probability can be made arbitrarily small for sufficiently large N as long as $R < I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma})$) and we use them on the M-block BF-AWGN channel, their expected word (or frame) error probability cannot exceed $P_{\text{out}}(R)$. In wireless speech transmission, frame error is a performance measure more sensible than bit error probability, because vocoders are frame-oriented [6].

In [7, 8] it is shown that $P_{\text{out}}(R)$ predicts quite accurately the performance of actual coding schemes for finite N (say $N \simeq 100$). Thus, outage-rate results have a practical relevance. □

In this paper we are interested in finding the power allocation strategies that minimize the outage rate under a constraint on the average transmitted power. To this purpose, we define the class of power allocation functions among which we search for the optimum.

Definition. The class of *statistical stationary memoryless power allocation functions* is the set of all time-invariant (measurable) random functions $\mathbb{R}_+^M \to \mathbb{R}_+^M$ (we denote the set of non-negative reals by $\mathbb{R}_+$). Thus, $\boldsymbol{\gamma}$ is an M-dimensional random vector with conditional cumulative distribution function (cdf) $F(\boldsymbol{\gamma} \mid \boldsymbol{\alpha})$[2]. Consider a sequence of M-block frames, indexed by $k = 0, 1, \ldots$. Let $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ be the channel gain and the power allocation vectors in the k-th frame. Then, the memoryless property of power allocation is expressed by:

$$F(\boldsymbol{\gamma}^{(0)}, \cdots, \boldsymbol{\gamma}^{(n-1)} | \boldsymbol{\alpha}^{(0)}, \cdots, \boldsymbol{\alpha}^{(n-1)}) = \prod_{k=0}^{n-1} F(\boldsymbol{\gamma}^{(k)} | \boldsymbol{\alpha}^{(k)})$$

---

[1] More generallly, we might consider an MN-dimensional power allocation $\boldsymbol{\gamma} = \{\gamma[k] : k = 0, \ldots, MN-1\}$. However, a simple application of Jensen's inequality shows that for any power allocation there exists a piecewise-constant power allocation satisfying the same average transmitted power constraints and yielding a larger instantaneous mutual information [15]. Thus, in the following we consider only piecewise-constant power-allocation functions.

[2] Given two random vectors $\mathbf{x}$ and $\mathbf{y}$, we adopt the somewhat improper but simple notation $F(\mathbf{y} \mid \mathbf{x})$ to indicate the more complete but cumbersome $F_{\mathbf{y}|\mathbf{x}}(\mathbf{v} \mid \mathbf{u}) \triangleq P(y_0 \le v_0, \ldots, y_{M-1} \le v_{M-1} \mid x_0 = u_0, \ldots, x_{M-1} = u_{M-1})$.

For the M-block channel we can relax the assumption that $\boldsymbol{\alpha} = \{\alpha_m\}_{m=0}^{M-1}$ is an M-vector taken from an ergodic process, since for finite M this would be irrelevant in any case. On the other hand, when we consider an infinite sequence of frames $(\ldots, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\alpha}^{(k)}, \ldots)$, we do assume that the M-vector process $\{\boldsymbol{\alpha}^{(k)}\}$ is ergodic (with respect to the index k, which can be regarded as a "time" index). We define the *long-term* average transmitted power as the limit of the normalized power sum over the whole sequence of frames, i.e.,

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathrm{E}[|x[kMN + mN + n]|^2 \mid \boldsymbol{\gamma}^{(k)}] = \lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \left\langle \boldsymbol{\gamma}^{(k)} \right\rangle \qquad (7)$$

(for a vector $\mathbf{a}$ of length M, we introduce the shorthand notation $\langle \mathbf{a} \rangle = \frac{1}{M} \sum_{m=0}^{M-1} a_m$). Under the assumptions of stationary and memoryless power allocation and of ergodic fading, (7) converges in probability to an ensemble average, so that we consider the power constraint $\mathrm{E}[\langle \boldsymbol{\gamma} \rangle] \leq \mathcal{P}$.

In [12], the notion of delay-limited capacity is introduced in the general case of discrete-time ergodic fading and multiple-access channel. In our context, the delay-limited capacity of the M-blocks BF-AWGN channel can be defined as

$$\mathcal{C}_{\text{delay}}(\mathcal{P}) = \max_{\boldsymbol{\gamma}: E[\langle \boldsymbol{\gamma} \rangle] \leq \mathcal{P}} \min_{\boldsymbol{\alpha}} I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \qquad (8)$$

In essence, the delay-limited capacity is the maximum rate at which one can transmit reliably for any fading realization, subject to the long-term power constraint. Equivalently, let $\widehat{P}_{\text{out}}(R, \mathcal{P})$ be the minimum outage rate subject to $E[\langle \boldsymbol{\gamma} \rangle] \leq \mathcal{P}$. Then, $\mathcal{C}_{\text{delay}}(\mathcal{P})$ is given by

$$\mathcal{C}_{\text{delay}}(\mathcal{P}) = \sup \left\{ R : \widehat{P}_{\text{out}}(R, \mathcal{P}) = 0 \right\} \qquad (9)$$

It should be clear that the minimum outage rate and the delay-limited capacity problems are closely related, and that the solution of the latter follows as a special case from the solution of the former.

III OPTIMAL POWER ALLOCATION

The constrained minimization problem can be stated as

$$\left\{ \begin{array}{ll} \text{Minimize} & P(I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma}) < R) \\ \text{Subject to} & E[\langle \boldsymbol{\gamma} \rangle] \leq \mathcal{P} \end{array} \right. \qquad (10)$$

The following proposition yields the general form of the solution:

Proposition 2. The problem (10) is solved by $\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ in the form

$$\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) = \left\{ \begin{array}{ll} \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) & \text{with probability } \widehat{w}(\boldsymbol{\alpha}) \\ \mathbf{0} & \text{with probability } 1 - \widehat{w}(\boldsymbol{\alpha}) \end{array} \right. \qquad (11)$$

where $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ is the solution of the problem

$$\left\{ \begin{array}{ll} \text{Minimize} & \langle \boldsymbol{\gamma} \rangle \\ \text{Subject to} & I_M(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \geq R \end{array} \right. \qquad (12)$$

and where $\widehat{w}(\boldsymbol{\alpha})$ is a weight function $\mathbb{R}_+^M \to [0, 1]$, solution of the problem

$$\left\{ \begin{array}{ll} \text{Maximize} & \mathrm{E}[w(\boldsymbol{\alpha})] \\ \text{Subject to} & 0 \leq w(\boldsymbol{\alpha}) \leq 1 \text{ and } \mathrm{E}[\langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle \, w(\boldsymbol{\alpha})] = \mathcal{P} \end{array} \right. \qquad (13)$$

The resulting minimum outage rate is

$$\widehat{P}_{\text{out}}(R, \mathcal{P}) = 1 - \mathrm{E}[\widehat{w}(\boldsymbol{\alpha})] \qquad (14)$$

□

In the following, we denote by $\mathcal{Q}$ the region of $\mathbb{R}_+^M$ where the components of $\boldsymbol{\alpha}$ are non-decreasing, i.e., $\mathcal{Q} = \{\boldsymbol{\alpha} \in \mathbb{R}_+^M : \alpha_0 \geq \cdots \geq \alpha_{M-1}\}$. The explicit form of $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ is given by next proposition and corollary:

Proposition 3. The vector $\widehat{\gamma}(\boldsymbol{\alpha})$ solving (12) has components given by

$$\widehat{\gamma}_m(\boldsymbol{\alpha}) = \left[\lambda(\boldsymbol{\alpha}) - \frac{1}{\alpha_m}\right]_+ \quad \text{for } m = 0, \dots, M-1 \tag{15}$$

where $\lambda(\boldsymbol{\alpha})$, obtained by solving the equality constraint $I_M(\boldsymbol{\alpha}, \widehat{\gamma}) = R$, is given by

$$\lambda(\boldsymbol{\alpha}) = \left(\frac{e^{2MR}}{\prod_{\ell=0}^{\mu-1} \alpha_\ell}\right)^{1/\mu} \tag{16}$$

and where, without loss of generality, we assume $\boldsymbol{\alpha} \in \mathcal{Q}$. Moreover, μ is the unique integer in $\{1, \dots, M\}$ such that $1/\alpha_m \leq \lambda(\boldsymbol{\alpha})$ for $m < \mu$ and $1/\alpha_m > \lambda(\boldsymbol{\alpha})$ for $m \geq \mu$. □

If $\boldsymbol{\alpha} \notin \mathcal{Q}$, it is sufficient to sort $\boldsymbol{\alpha}$ in non-increasing order. Let π denote the sorting permutation and $\pi(\boldsymbol{\alpha})$ the sorted fading vector. Then, the power allocation $\widehat{\gamma}$ is obtained by permuting the indices of (15) according to π and by using $\pi(\boldsymbol{\alpha})$ instead of $\boldsymbol{\alpha}$ in (16).

Corollary 1. $\widehat{\gamma}(\boldsymbol{\alpha})$ given in (15) is continuous for $\boldsymbol{\alpha} \in \mathbb{R}_+^M, \boldsymbol{\alpha} \neq \mathbf{0}$ and $\langle \widehat{\gamma}(\boldsymbol{\alpha}) \rangle$ is a non-increasing function of α_m, for all $0 \leq m \leq M-1$. □

In order to characterize the optimal power allocation $\widetilde{\gamma}(\boldsymbol{\alpha})$ of Proposition 2 we finally need to find $\widehat{w}(\boldsymbol{\alpha})$ solution of (13). We start with the following lemma:

Lemma 1. Let u be a non-negative real random variable with cdf $F(u)$. The solution of the problem

$$\left\{\begin{array}{ll} \text{Maximize} & \mathrm{E}[w(u)] \\ \text{Subject to} & 0 \leq w(u) \leq 1 \text{ and } \mathrm{E}[u\, w(u)] = \mathcal{P} \end{array}\right. \tag{17}$$

is always in the form

$$\widehat{w}(u) = \left\{\begin{array}{ll} 1 & \text{for } u < s^* \\ w_0 & \text{for } u = s^* \\ 0 & \text{for } u > s^* \end{array}\right. \tag{18}$$

where $w_0 \in [0,1]$ and $s^* \geq 0$ depend on $F(u)$ and on $\mathcal{P}$ and are given as follows. Define $\mathcal{P}(s)$ and $\overline{\mathcal{P}}(s)$ by

$$\mathcal{P}(s) = \int_{[0,s)} u\, dF(u)\ , \quad \overline{\mathcal{P}}(s) = \int_{[0,s]} u\, dF(u) \tag{19}$$

Then, $s^* = \sup\{s : \mathcal{P}(s) < \mathcal{P}\}$ and

$$w_0 = \frac{\mathcal{P} - \mathcal{P}(s^*)}{\overline{\mathcal{P}}(s^*) - \mathcal{P}(s^*)} \tag{20}$$

□

In passing, we note that $w_0 = 0$ if $\mathcal{P} = \mathcal{P}(s^*)$ and $w_0 \to 1$ if $\mathcal{P} \to \overline{\mathcal{P}}(s^*)$. Moreover, if $F(u)$ is continuous in $u = s^*$, then $\overline{\mathcal{P}}(s^*) - \mathcal{P}(s^*) = s^* P(u = s^*) = 0$, so that $\{u = s^*\}$ is a set of probability measure zero and the value of w_0 can be any real in $[0,1]$ without affecting neither the objective neither the constraint of (17). The function $\widehat{w}(u)$ defined above satisfies the constraint with equality,

$$\begin{aligned} E[u\, \widehat{w}(u)] &= \mathcal{P}(s^*) + w_0 s^* P(u = s^*) \\ &= \mathcal{P}(s^*) + \frac{\mathcal{P} - \mathcal{P}(s^*)}{\overline{\mathcal{P}}(s^*) - \mathcal{P}(s^*)} s^* P(u = s^*) \\ &= \mathcal{P} \end{aligned}$$

We are now ready to state:

Proposition 4. The solution $\widehat{w}(\boldsymbol{\alpha})$ of (13) is given by

$$\widehat{w}(\boldsymbol{\alpha}) = \begin{cases} 0 & \text{for } \boldsymbol{\alpha} \notin \overline{\mathcal{R}}(s^*) \\ w_0 & \text{for } \boldsymbol{\alpha} \in \mathcal{B}(s^*) \\ 1 & \text{for } \boldsymbol{\alpha} \in \mathcal{R}(s^*) \end{cases} \tag{21}$$

where:

1. For $s \in \mathbb{R}_+$ we define the region $\mathcal{R}(s) = \{\boldsymbol{\alpha} \in \mathbb{R}_+^M : \langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle < s\}$ and its closure $\overline{\mathcal{R}}(s) = \{\boldsymbol{\alpha} \in \mathbb{R}_+^M : \langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle \leq s\}$.
2. The *boundary surface* $\mathcal{B}(s)$ of $\overline{\mathcal{R}}(s)$ is the set of points $\boldsymbol{\alpha}$ such that $\langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle = s$.
3. The two average power sums $\mathcal{P}(s)$ and $\overline{\mathcal{P}}(s)$ are defined by

$$\mathcal{P}(s) = \int_{\mathcal{R}(s)} \langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle \, dF(\boldsymbol{\alpha}) \quad \text{and} \quad \overline{\mathcal{P}}(s) = \int_{\overline{\mathcal{R}}(s)} \langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle \, dF(\boldsymbol{\alpha}) \tag{22}$$

4. The power sum threshold s^* is defined by $s^* = \sup\{s : \mathcal{P}(s) < \mathcal{P}\}$.
5. The value w_0 in (21) is given by

$$w_0 = \frac{\mathcal{P} - \mathcal{P}(s^*)}{\overline{\mathcal{P}}(s^*) - \mathcal{P}(s^*)} \tag{23}$$

Comment. The proposition follows from Lemma 1 and Corollary 1. First, note that the power sum $u = \langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle$ is a function $\mathbb{R}_+^M \to \mathbb{R}_+$ and for a given $F(\boldsymbol{\alpha})$, it is a non-negative random variable with cdf

$$F(u) = P(\langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle \leq u) = \int_{\overline{\mathcal{R}}(u)} dF(\boldsymbol{\alpha})$$

By applying Lemma 1 to u, we find s^* and w_0 which maximize $E[w(u)]$ subject to the power constraint, that can be rewritten as $E[u\, w(u)] \leq \mathcal{P}$. Then, from Corollary 1 we have that $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ is continuous and that $\langle \widehat{\boldsymbol{\gamma}} \rangle$ has componentwise non-positive gradient. This implies that for any $0 \leq s < s' < \infty$, the regions $\mathcal{R}(s)$ and $\mathcal{R}(s')$ defined by Proposition 4 satisfy $\mathcal{R}(s) \subseteq \mathcal{R}(s')$. Then, the intervals $u < s^*$, $u = s^*$ and $u > s^*$ on the non-negative u-axis correspond to the regions $\mathcal{R}(s^*)$, to the boundary surface $\mathcal{B}(s^*)$ and to the complement region $\mathcal{U}(s^*) = \mathbb{R}_+^M - \overline{\mathcal{R}}(s^*)$, respectively.

The operational significance of Proposition 4 is the following. The weight function (21) corresponds to setting a threshold s^* so that, if the power sum $\langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle$ exceeds this threshold, transmission is turned off, while if $\langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle$ is below the threshold, transmission is turned on and the power is allocated according to $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha})$, i.e., the allocation that requires the minimum power sum in order to avoid an outage event (see (12)). The threshold s^* is chosen so that the long-term power is actually equal to $\mathcal{P}$. If $F(\boldsymbol{\alpha})$ has discontinuities, a randomization may be needed if $\langle \widehat{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \rangle$ is exactly equal to s^*, i.e., if $\boldsymbol{\alpha} \in \mathcal{B}(s^*)$. Again, the probability w_0 of transmitting in this case is chosen so that the long-term power is equal to $\mathcal{P}$.

The optimal power allocation policy has an *economic interpretation.* If fading is very bad, compensating for it would cost too much in terms of average power. Hence, it is better to accept an outage and turn off transmission, so as to we save power for less unfavorable fading conditions. □

From (14), the resulting minimum outage rate can be written as

$$\widehat{P}_{\text{out}}(R, \mathcal{P}) = 1 - w_0 P(\boldsymbol{\alpha} \in \mathcal{B}(s^*)) - P(\boldsymbol{\alpha} \in \mathcal{R}(s^*)) \tag{24}$$

In the case of continuous fading (e.g., Rayleigh, Rice, Log-Normal and Nakagami fading have a continuous cdf [9]), the boundary surface $\mathcal{B}(s^*)$ is a set of probability measure zero. Then, the outage rate is simply given by the integral $\widehat{P}_{\text{out}} = \int_{\mathcal{U}(s^*)} f(\boldsymbol{\alpha})\, d\boldsymbol{\alpha}$ where $f(\boldsymbol{\alpha})$ is the fading pdf. We will refer to the region $\mathcal{U}(s^*)$ as the *outage region.* We can obtain an explicit characterization of $\mathcal{U}(s^*)$ as follows. First, we partition the region $\mathcal{Q} \subset \mathbb{R}_+^M$ into M subregions $\mathcal{W}_i$, for $i = 1, \ldots, M$, defined by the inequalities

$$\left\{ \begin{array}{ll} \left(\frac{e^{2MR}}{\prod_{\ell=0}^{i-1} \alpha_\ell} \right)^{1/i} \geq \frac{1}{\alpha_m} & m = 0, \ldots, i-1 \\ \left(\frac{e^{2MR}}{\prod_{\ell=0}^{i-1} \alpha_\ell} \right)^{1/i} < \frac{1}{\alpha_m} & m = i, \ldots, M-1 \end{array} \right. \tag{25}$$

(note that, $\{\mathcal{W}_i : i = 1, \ldots, M\}$ is a partition of $\mathcal{Q}$ because of the existence and uniqueness of μ in (16)). Then, the intersection of $\mathcal{U}(s^*)$ with $\mathcal{W}_i$, for all $i = 1, \ldots, M$, is given by

$$\mathcal{U}(s^*) \cap \mathcal{W}_i = \left\{ \boldsymbol{\alpha} \in \mathbb{R}_+^M : \frac{i}{M} \left[\left(\frac{e^{2MR}}{\prod_{\ell=0}^{i-1} \alpha_\ell} \right)^{1/i} - \frac{1}{i} \sum_{m=0}^{i-1} \frac{1}{\alpha_m} \right] > s^* \right\} \tag{26}$$

By permuting the components of $\boldsymbol{\alpha}$ according to all possible permutations in both (25) and (26), we obtain all the inequalities defining $\mathcal{U}(s^*)$ (see the two-dimensional example in Section IV).

General solution for the delay-limited capacity. From Proposition 4 we see that the zero-outage condition is achieved for $s^* \to \infty$, i.e., when transmission is never turned off. Then, the delay-limited capacity can be computed from the limit $\mathcal{P}_\infty = \lim_{s \to \infty} \mathcal{P}(s)$. If this is infinite, then $\mathcal{C}_{\text{delay}}(\mathcal{P}) = 0$, since we would need infinite average power to transmit reliably at rate $R > 0$. If, on the contrary, $\mathcal{P}_\infty$ is finite, this is the average power needed for reliable transmission at rate R for all fading states. For a given M and fading cdf $F(\boldsymbol{\alpha})$, $\mathcal{P}_\infty$ is a function of R, so that we can write $\mathcal{P} = \mathcal{P}_\infty(R)$. By inverting this functional relation we obtain the delay-limited capacity curve $R = \mathcal{C}_{\text{delay}}(\mathcal{P})$.

Blind and adaptive power allocation. The optimal power allocation policy $\widetilde{\gamma}(\boldsymbol{\alpha})$ depends on the fading statistics only through the threshold s^*. In actual implementations, the fading statistics may be unknown. Then, algorithms for blind estimation of s^* should be devised. A simple heuristic algorithm is the following: i) Initialize $s_0^* = \mathcal{P}$. ii) For $k = 1, 2, \ldots$, let

$$s_k^* = s_{k-1}^* \left[1 + \epsilon \left(\mathcal{P} - \frac{1}{k} \sum_{i=0}^{k-1} \left\langle \widetilde{\gamma}(\boldsymbol{\alpha}^{(i)}) \right\rangle \right) \right] \tag{27}$$

and use the threshold s_k^* to determine $\widetilde{\gamma}(\boldsymbol{\alpha}^{(k)})$.

Note that $\frac{1}{k} \sum_{i=0}^{k-1} \left\langle \widetilde{\gamma}(\boldsymbol{\alpha}^{(i)}) \right\rangle$ is the actual transmitted long-term power. Then, if this is less than $\mathcal{P}$, the threshold is increased, while if this is larger than $\mathcal{P}$, the threshold is decreased. As usual, the stepsize must be chosen by trading-off acquisition time and tracking precision. The use of two different stepsizes, one for acquisition mode and the other for tracking mode, might be a good option. In order to track a slowly-varying fading statistics, the algorithm can be made adaptive by using an exponentially decaying window with an appropriate forgetting factor.

IV APPLICATIONS TO RAYLEIGH FADING CHANNELS

Consider a BF-AWGN channel made by a set of M independent Rayleigh fading channels. The gain of the m-th channel is exponentially distributed as $\alpha_m \sim f(\alpha_m) = e^{-\alpha_m}$ [3]. For $M = 1$ and constant power transmission, the outage rate is given by $\widehat{P}_{\text{out}}(R, \mathcal{P}) = 1 - \exp(-(e^{2R} - 1)/\mathcal{P})$ [2]. With optimal power allocation s^* is the solution of $\mathcal{P}(s) = \mathcal{P}$, with $\mathcal{P}(s)$ given by $\mathcal{P}(s) = (e^{2R} - 1)\mathrm{Ei}\,(1, \frac{\mathrm{e}^{2\mathrm{R}} - 1}{\mathrm{s}})$ (where $\mathrm{Ei}\,(\mathrm{n}, \mathrm{x}) = \int_1^\infty \mathrm{t}^{-\mathrm{n}} \mathrm{e}^{-\mathrm{xt}} \mathrm{dt}$, for $\mathrm{Re}\{x\} > 0$) and the resulting outage rate is $\widehat{P}_{\text{out}}(R, \mathcal{P}) = 1 - \exp(-(e^{2R} - 1)/s^*)$. Fig. 1 shows the above outage rates vs. $\mathcal{P}$ for

$R = 1.0 \log 2$ nat/symbol. We observe that optimal power allocation yields a very large power gain in terms of outage rate with respect to constant power transmission (about 22 dB at $P_{\text{out}} = 10^{-3}$ for $R = 0.5 \log 2$). This behavior differs remarkably from that of average capacity, where optimal power allocation yields only a very small gain (see [9, 10]). If $M = 1$ (no interleaving) the delay-limited capacity is zero since $\mathcal{P}_\infty = \lim_{s\to\infty}(e^{2R}-1)\text{Ei}\,(1, \frac{e^{2R}-1}{s}) = \infty$.

Fig. 2 shows $\mathcal{P}(s)$ vs. s for $R = 0.5 \log 2$ nat/symbol. For comparison, the outage rate for constant power allocation in the general case of M independent blocks is immediately calculated from the M-fold convolution of the distribution of $\log(1+\alpha\mathcal{P})$ [2]. Fig. 3 shows the above outage rates vs. $\mathcal{P}$ for $1.0 \log 2$ nat/symbol. The power gain obtained by optimal power allocation with respect to constant power is still very large, especially for large R, but it is reduced with respect to the case $M = 1$ (e.g., for $R = 0.5 \log 2$ it is about 15.5 dB at $P_{\text{out}} = 10^{-3}$). This fact can be explained by observing that, as $M \to \infty$, both outage rates for optimal and for constant power allocations tend to their average capacity limits, i.e., tend to be large if $R > \mathcal{C}_{\text{pow}}(\mathcal{P})$ (resp., if $R > \mathcal{C}_{\text{constr}}(\mathcal{P})$) and vanishing if $R < \mathcal{C}_{\text{pow}}(\mathcal{P})$ (resp., if $R < \mathcal{C}_{\text{constr}}(\mathcal{P})$). Since $\mathcal{C}_{\text{pow}}(\mathcal{P})$ and $\mathcal{C}_{\text{constr}}(\mathcal{P})$ are very close, we expect that the gain of optimal with respect to constant power allocation reduces as M increases.

The optimal power allocation yields zero outage rate for $\mathcal{P}$ larger than a threshold (denoted by the vertical solid lines in the figures). This power threshold is the delay-limited capacity limit $\mathcal{P} = \mathcal{P}_\infty(R)$. We can compute explicitly $\mathcal{P}_\infty$ by using the asymptotic expansion $\text{Ei}\,(1, z) = -\Gamma - \log z + O(z)$ for $z \to 0$. We obtain:

$$\lim_{s\to\infty} \mathcal{P}(s) = \mathcal{P}_\infty(R) = (e^{4R}-1)\log\left(\frac{2\sinh(4R)}{e^{4R}-1}\right) + e^{2R}(\pi - 4\arctan(e^{-2R})) - 4R \qquad (28)$$

From the delay-limited capacity curve $R = \mathcal{C}_{\text{delay}}(\mathcal{P})$ obtained from the above limit we can see that the 2-block Rayleigh BF channel is very close to the AWGN capacity for all rates (about 5 dB power penalty with respect to AWGN). We did not attempt closed-form evaluation of the delay-limited capacity for $M > 2$: however, this can be computed by Monte Carlo simulation of $\mathcal{P}(s)$ for very large s and given R.

Finally, Fig. 4 shows the long-term average transmitted power when the algorithm (27) is used to estimate the optimal threshold s^*. This curve was obtained by Monte Carlo simulation with $\mathcal{P} = 2$ dB, $M = 2$ and $R = 0.5 \log 2$ nat/symbol. We used a constant step size $\epsilon = 0.01$. The algorithm is able to approach the desired long-term power even if the fading statistics is totally unknown a priori.

V CONCLUSIONS

In this paper we discussed the relevance of the outage rate criterion (also referred to as "outage capacity" [4]) for wireless real-time speech transmission. We solved the minimum outage rate problem for the BF-AWGN channel with transmission over M parallel blocks and perfect transmitter and receiver CSI. The optimal power allocation minimizing the outage rate is independent of the fading statistics and correlation, except for the value of the power threshold above which transmission must be turned off. We presented a blind adaptive algorithm that estimates the optimal value of the power threshold without prior knowledge of the fading statistics and correlation. We also showed that the minimum outage rate problem is closely related to the delay-limited capacity problem and that our solution provides a tool to compute the delay-limited capacity of the M-block BF-AWGN channel. This method is general, although closed form results appear to be difficult to obtain for correlated fading and/or $M > 2$. Nevertheless, Monte Carlo simulation can be used to compute both the minimum outage rate and the delay-limited capacity in the general case. We also derived closed-form results for the independent Rayleigh fading channel with $M = 1$ and $M = 2$ blocks.

Impressive power gains can be achieved by optimal power allocation with respect to constant power transmission. Moreover, reliable communication is possible with a transmission delay $M = 2$ and with only 5 dB penalty with respect to AWGN (for $M = 4$ the power penalty reduces to about 3 dB). These results show that optimal power allocation has a dramatic impact on the performance of a delay-limited system, while for delay-unconstrained systems (i.e., in terms of average capacity) its benefit is very reduced (see [9, 10]). From a practical viewpoint, multicarrier coded transmission with optimal power allocation and a small number M of sufficiently spaced carriers emerges as a practically viable technique for high-performance wireless systems with delay constraints. This technique provides large improvements especially in the case of slow fading, typical of personal communication systems where mobile terminals move at walking speed. Future work will consider the performance and the construction of actual codes with optimal power allocation, and the combination of optimal power allocation with transmission antenna diversity [16].

REFERENCES

[1] J. Proakis, *Digital Communications.* 3rd Edition. New York: McGraw-Hill, 1995.

[2] L. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Vehic. Tech.,* Vol. 43, No. 2, May 1994.

[3] G. Kaplan and S. Shamai (Shitz), "Error probabilities for the block-fading Gaussian channel," *A.E.Ü.,* Vol. 49, No. 4, pp. 192 - 205, 1995.

[4] G. Foschini Jr. and M. Gans, "On limits of wireless communication in a fading environment when using multiple antennas," *Wireless Personal Communications,* submitted for publication, 1997.

[5] R. McEliece and W. Stark, "Channels with block interference," *IEEE Trans. Inform. Theory,* Vol. 30, No. 1, pp. 44 - 53, Jan. 1984.

[6] T. Rappaport, *Wireless Communications.* Englewood Cliffs, NJ: Prentice-Hall, 1996.

[7] R. Knopp, P. A. Humblet, "On coding for block-fading channels," submitted to *IEEE Trans. on Commun.* March 1997.

[8] R. Knopp, P. A. Humblet, "Maximizing diversity on block fading channels," *Proceedings of IEEE ICC '97,* Montréal, Canada, 8–12 June, 1997.

[9] A. Goldsmith, *Design and Performance of High-Speed Communication Systems over Time-Varying Radio Channels.* Ph.D. Thesis, University of California at Berkeley, Berkeley, CA, USA, 1994. Also: UCB/ERL Memorandum M94/75.

[10] A. Goldsmith and P. Varaiya, "Increasing spectral efficiency through power control," *Proceedings of IEEE ICC '93*, pp. 600–604, Geneva, Switzerland, June 1993.

[11] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," to appear on *IEEE Trans. on Inform. Theory,* 1997.

[12] D. Tse and V. Hanly, "Multi-access fading channels: Part II, Delay-limited capacities," submitted to *IEEE Trans. on Inform. Theory,* 1996.

[13] J. Wolfowitz, *Coding Theorems of Information Theory.* New York: Springer-Verlag, 1978.

[14] G. Caire, R. Knopp and P. Humblet, "System capacity of F-TDMA cellular systems," submitted to *IEEE Trans. Commun.,* April 1997.

[15] G. Caire, G. Taricco and E. Biglieri, "Information outage rate and power control," in preparation, to be submitted to *IEEE Trans. on Inform. Theory,* August 1997.

[16] V. Tarokh, N. Seshadri and R. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criteria," proceedings of *ICC '97*, Montréal, Canada, 8 - 12 June, 1997.

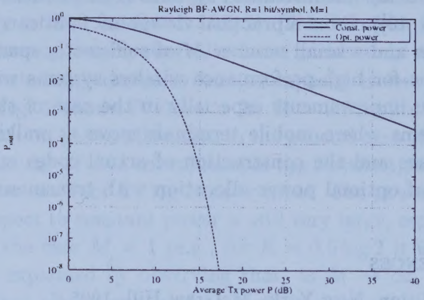


Figure 1: Outage rate vs $\mathcal{P}$ for the Rayleigh fading channel with $M = 1$ and $R = 1.0 \log 2$ nat/symbol.

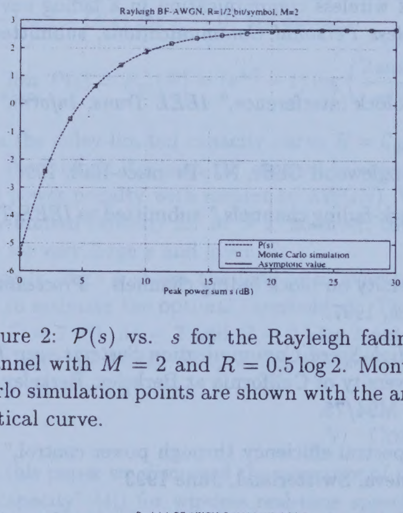


Figure 2: $\mathcal{P}(s)$ vs. s for the Rayleigh fading channel with $M = 2$ and $R = 0.5 \log 2$. Monte Carlo simulation points are shown with the analytical curve.

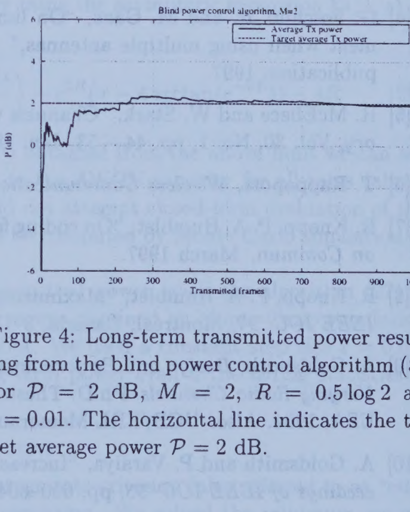


Figure 4: Long-term transmitted power resulting from the blind power control algorithm (55) for $\mathcal{P} = 2$ dB, $M = 2$, $R = 0.5 \log 2$ and $\epsilon = 0.01$. The horizontal line indicates the target average power $\mathcal{P} = 2$ dB.

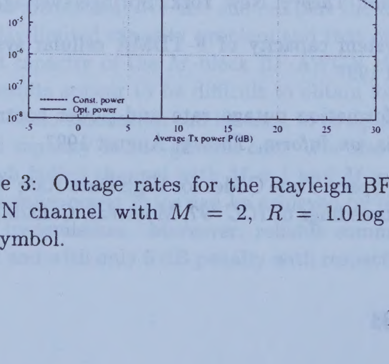


Figure 3: Outage rates for the Rayleigh BF-AWGN channel with $M = 2$, $R = 1.0 \log 2$ nat/symbol.

Multiple-Antenna Communications When Nobody Knows the Rayleigh Fading Coefficients

THOMAS L. MARZETTA BERTRAND M. HOCHWALD

Bell Laboratories
Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974
{tlm, hochwald}@bell-labs.com

Abstract

We analyze a mobile wireless link comprising M transmitter and N receiver antennas operating in a Rayleigh flat-fading environment. The propagation coefficients between every pair of transmitter and receiver antennas are statistically independent and unknown; they remain constant for a coherence interval of T symbol periods, after which they change to new independent values which they maintain for another T symbol periods, and so on.

Computing the link capacity, associated with channel coding over multiple fading intervals, requires an optimization over the joint density of $T \cdot M$ complex transmitted signals. We prove that there is no point in making the number of transmitter antennas greater than the length of the coherence interval: the capacity for $M > T$ is equal to the capacity for $M = T$. Capacity is achieved when the $T \times M$ transmitted signal matrix is equal to the product of two statistically independent matrices: a $T \times T$ isotropically distributed unitary matrix times a certain $T \times M$ random matrix that is diagonal, real, and nonnegative. This result enables us to determine capacity for many interesting cases. We conclude that, for a fixed number of antennas, as the length of the coherence interval increases, the capacity approaches the capacity obtained as if the receiver knew the propagation coefficients.

1 Introduction

It is likely that future breakthroughs in wireless communication will be driven largely by high data rate applications. Sending video rather than speech, for example, increases the data rate by two or three orders-of-magnitude. Increasing the link or channel bandwidth is a simple but costly—and ultimately unsatisfactory—remedy. A more economical solution is to exploit propagation diversity through multiple-element transmitter and receiver antenna arrays.

It has been shown [1, 5] that, in a Rayleigh flat-fading environment, a link comprising multiple-element antennas has a theoretical capacity that increases linearly with the smaller of the number of transmitter and receiver antennas, provided that the complex-valued propagation coefficients between all pairs of transmitter and receiver antennas are statistically independent and known to the receiver (but not the transmitter). If the time between signal fades is sufficiently long—often a reasonable assumption for a fixed wireless environment—then the transmitter can send training signals that allow the receiver to estimate the propagation coefficients accurately, and the results of [1, 5] are applicable.

With a mobile receiver, however, the time between fades may be too short to permit reliable estimation of the coefficients. A 60 mile/hour mobile operating at 1.9 GHz has a fading interval of about 3 ms, which for a bandwidth of 30 kHz, corresponds to only about 100 symbol periods. We approach the problem of determining the capacity of a time-varying multiple-antenna communication channel, using the tools of information theory, and without any ad hoc training schemes in mind.

Our model for the link assumes M transmitter and N receiver antennas. The propagation coefficients, which neither the transmitter nor the receiver knows, are constant for T symbol periods, after which they change to new independent random values which they maintain for another T symbol periods, and so on. To a first order approximation, this models a time-varying or frequency-hopping fading channel (see also Jakes' model [3]), where T represents the coherence interval of the fades or the time between frequency hops. The random propagation coefficients are modelled as independent, identically distributed, zero-mean, circularly symmetric complex Gaussian random variables. Thus there are two sources of noise at work: multiplicative noise that is associated with the Rayleigh fading, and the usual additive receiver noise. The link is completely described by the conditional probability density of the $T \cdot N$ complex received signals given the $T \cdot M$ complex transmitted signals. Although this conditional density is complex Gaussian, the transmitted signals affect only the conditional covariance (rather than the mean) of the received signals—a source of difficulty in the problem.

If one performs channel coding over multiple independent fading intervals, information theory tells us that it is theoretically possible to transmit information reliably at a rate that is bounded by the channel capacity. Computing the capacity involves finding the joint probability density function of the TM-dimensional transmitted signal that maximizes the mutual information between it and the TN-dimensional received signal. The special case $M = N = T = 1$ is addressed in [6], where it is shown that the maximizing transmitted signal density is discrete and has support only on the nonnegative real axis. The maximization appears, in general, to be computationally intractable for $M > 1$ or $T > 1$.

Nevertheless, we show that the dimensionality of the maximization can be reduced from $T \cdot M$ to $\min(M, T)$, and that the capacity can therefore be easily computed for many nontrivial cases. In the process, we determine the signaling schemes that achieve capacity and find the asymptotic dependence of the capacity on T.

2 Multiple-Antenna Link

Figure 1 displays a communication link or channel comprising M transmitter antennas and N receiver antennas that operates in a Rayleigh flat-fading environment. Each receiver antenna responds to each transmitter antenna through a statistically independent fading coefficient that is constant for T symbol periods. The fading coefficients are not known by either the transmitter or the receiver. The received signals are corrupted by additive noise that is statistically independent among the N receivers and the T symbol

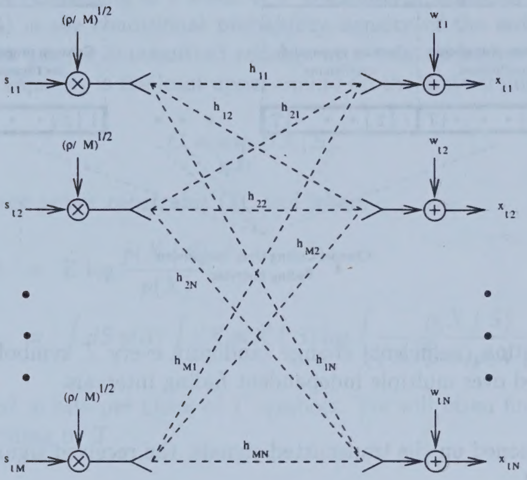


Figure 1: Wireless link comprising M transmitter and N receiver antennas. Every receiver antenna is connected to every transmitter antenna through an independent, random, unknown propagation coefficient having Rayleigh distributed magnitude and uniformly distributed phase. Normalization ensures that the total expected transmitted power is independent of M for a fixed ρ.

periods.

The complex-valued signal x_{tn} that is measured at receiver antenna n, and discrete time t, is given by

$$x_{tn} = \sqrt{\rho/M} \sum_{m=1}^{M} h_{mn} s_{tm} + w_{tn}, \qquad t = 1, \ldots T, \ \ n = 1 \ldots N. \tag{1}$$

Here h_{mn} is the complex-valued fading coefficient between the mth transmitter antenna and the nth receiver antenna. The fading coefficients are constant for $t = 1, \ldots T$, and they are independent and $\mathcal{CN}(0,1)$ (complex normal) distributed, with density

$$p(h_{mn}) = (1/\pi) \exp\left\{-|h_{mn}|^2\right\}.$$

The complex-valued signal that is fed at time t into transmitter antenna m is denoted s_{tm}, and its average (over the M antennas) expected power is equal to one. This may be written

$$\frac{1}{M} \sum_{m=1}^{M} \mathrm{E}\, |s_{tm}|^2 = 1, \qquad t = 1, \ldots, T. \tag{2}$$

The additive noise at time t and receiver antenna n is denoted w_{tn}, and is independent (with respect to both t and n), identically distributed $\mathcal{CN}(0,1)$. The quantities in the signal model (1) are normalized so that ρ represents the expected signal-to-noise ratio (SNR) at each receiver antenna, independently of M.

Both the fading coefficients and the receiver noise are complex Gaussian distributed.

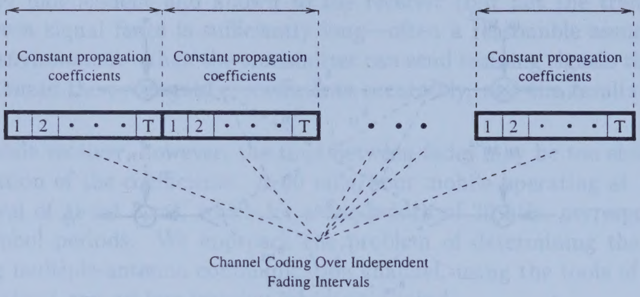

Constant propagation coefficients
Constant propagation coefficients
Constant propagation coefficients
1 2 · · · T
1 2 · · · T
1 2 · · · T
• • •
Channel Coding Over Independent Fading Intervals

Figure 2: Propagation coefficients change randomly every T symbol periods. Channel coding is performed over multiple independent fading intervals.

As a result, conditioned on the transmitted signals, the received signals are jointly complex Gaussian. Let

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1M} \\ \vdots & & \\ s_{T1} & \ldots & s_{TM} \end{bmatrix}, \qquad X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & & \\ x_{T1} & \ldots & x_{TN} \end{bmatrix},$$

$$H = \begin{bmatrix} h_{11} & \ldots & h_{1N} \\ \vdots & & \\ h_{M1} & \ldots & h_{MN} \end{bmatrix}, \qquad W = \begin{bmatrix} w_{11} & \ldots & w_{1N} \\ \vdots & & \\ w_{T1} & \ldots & w_{TN} \end{bmatrix},$$

where S is the $T \times M$ matrix of transmitted signals, X is the $T \times N$ matrix of received signals, H is the $M \times N$ matrix of propagation coefficients, and W is the $T \times N$ matrix of additive noise components. Then

$$X = \sqrt{\frac{\rho}{M}} SH + W. \tag{3}$$

It immediately follows that

$$p(X \mid S) = \frac{\exp\left(-\mathrm{tr}\left\{\left[I_T + (\rho/M)\, SS^\dagger\right]^{-1} XX^\dagger\right\}\right)}{\pi^{TN} \det^N \left[I_T + (\rho/M)\, SS^\dagger\right]}, \tag{4}$$

where I_T denotes the $T \times T$ identity matrix and "tr" denotes "trace."

The channel is completely described by this conditional probability density. Note that the propagation coefficients do not appear in this expression. Although the received signals are conditionally Gaussian, the transmitted signals only affect the covariance of the received signals, in contrast to the classical additive Gaussian noise channel where the transmitted signals affect the mean of the received signals.

We assume that the fading coefficients change to new independent realizations every T symbol periods. By performing channel coding over multiple fading intervals, as in Figure 2, the intervals of favorable fading compensate for the intervals of unfavorable fading.

Each channel use (consisting of a block of T transmitted symbols) is independent of every other, and (4) is the conditional probability density of the output X, given the input S. Thus, data can be transmitted reliably at a rate no higher than the channel capacity, where the capacity is the least upper bound on the mutual information between X and S, or

$$C = \sup_{p(S)} I(X;S),$$

subject to the average power constraint (2), and where

$$\begin{aligned} I(X;S) &= \mathrm{E} \log \frac{p(X \mid S)}{p(X)} \\ &= \int dS\, p(S) \int dX\, p(X \mid S) \log \left\{ \frac{p(X \mid S)}{\int d\acute{S}\, p(\acute{S}) p(X \mid \acute{S})} \right\}. \end{aligned} \tag{5}$$

Thus, C is measured in bits per block of T symbols. We will often find it convenient to normalize C by dividing by T.

3 Signals that Achieve Capacity

Direct analytical or numerical maximization of the mutual information in (5) is hopelessly difficult whenever $T \cdot M$, the number of components of the transmitted signal matrix S, is much greater than one. This section shows that the maximization effort can be reduced to a problem in $\min(M, T)$ dimensions, making it possible to compute capacity easily for many significant cases. Because of space limitations, we state the results without proofs and refer the interested reader to [2].

We observe that the effect of the transmitted signals on the conditional probability density of the received signals is through the $T \times T$ matrix $SS^\dagger$. It is therefore reasonable to expect that any possible joint probability density of the elements of $SS^\dagger$ can be realized with at most $M = T$ transmitter antennas.

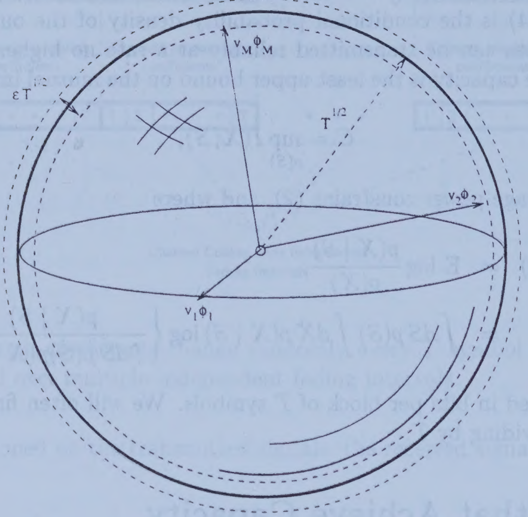
Theorem 1 *(Capacity for $M > T$ equals capacity for $M = T$) For any coherence interval T and any number of receiver antennas, the capacity obtained with $M > T$ transmitter antennas can be also be obtained by $M = T$ transmitter antennas.*

In what follows we assume that $M \leq T$.

Theorem 2 *(Structure of signal that achieves capacity) The signal matrix that achieves capacity can be written as $S = \Phi V$, where Φ is an $T \times T$ isotropically distributed unitary matrix, and V is an independent $T \times M$ real, nonnegative, diagonal matrix. Furthermore, we can choose the joint density of the diagonal elements of V to be unchanged by rearrangements of its arguments.*

An *isotropically distributed* unitary matrix has a probability density that is unchanged when the matrix is multiplied by any deterministic unitary matrix. In a natural way, an isotropically distributed unitary matrix is the $T \times T$ counterpart of a complex scalar having unit magnitude and uniformly distributed phase. More details, including the probability density of these matrices, may be found in [4] and [2].

Thus, the problem of maximizing $I(X;S)$ with respect to the joint probability density of the $T \cdot M$ complex elements of S reduces to the simpler problem of maximizing $I(X;S)$



T-dimensional complex sphere with radius $T^{1/2}$

Figure 3: The transmitted signals that achieve capacity are mutually orthogonal with respect to time. The constituent orthonormal unit vectors are isotropically distributed and independent of the signal magnitudes, which have mean-square value T. The solid sphere of radius $T^{1/2}$ demarcates the root mean-square. For $T \gg M$, the vectors all lie approximately on the surface of this sphere.

with respect to the joint probability density of the M nonnegative real diagonal elements of V. This joint probability can be constrained to be invariant to rearrangements of its arguments, and thus to obey $\mathrm{E}\, v_m^2 = T$ for all m. The mth column of S, representing the T complex signals that are fed into the mth transmitter antenna, is equal to the real nonnegative scalar v_m times an independent T-dimensional isotropically distributed complex unit vector ϕ_m. Here ϕ_m is the mth column of the $T \times T$ isotropically distributed unitary matrix Φ, and the M signal vectors $v_1\phi_1, \ldots, v_M\phi_M$ are mutually orthogonal. Figure 3 shows the signal vectors associated with the M transmitter antennas. Each signal vector is a T-dimensional complex vector (comprising $2T$ real components). The solid sphere demarcates the root mean-square values of the vector lengths; that is, $\mathrm{E}\, v_m^2 = T$. Later we argue that, for $T \gg M$, the magnitudes of the signal vectors are approximately $\sqrt{T}$ with very high probability.

4 Capacity and Capacity Bounds

In this section, we calculate capacity and some upper and lower bounds in some simple but nontrivial cases.

4.1 Capacity upper bound

An upper bound is obtained if we assume that the receiver is provided with a noise-free measurement of the propagation coefficients H. This *perfect-knowledge* upper bound is

$$C_u = T \cdot \mathrm{E} \log\det\left[I_N + \frac{\rho}{M} H^\dagger H\right], \tag{6}$$

and is derived in [2]. Equation (6) gives the upper bound per block of T symbols. The normalized bound, C_u/T, is independent of T. When H is known to the receiver, the perfect-knowledge capacity bound is achieved with transmitted signals S that are independent $\mathcal{CN}(0,1)$ (see also [5]).

In our model H is unknown to the receiver, but we intuitively expect C to approach C_u as T becomes large because a small portion of the coherence interval can be reserved for sending training data from which the receivers can estimate the propagation coefficients. When H is unknown and T is large, we therefore also expect the joint probability density of the capacity-achieving transmitted signals S to become approximately independent $\mathcal{CN}(0,1)$. We reconcile this intuition with the structure that is demanded in Theorem 2,

$$S = \Phi V = \left[\begin{array}{cccc} v_1\phi_1 & v_2\phi_2 & \dots & v_M\phi_M \end{array}\right],$$

where $\{\phi_m\}_{m=1}^M$ are the column vectors of Φ, by observing that, for fixed M, as T grows large two interesting things happen. First, the M complex random orthogonal unit vectors that comprise S become virtually independent. Second, for any $\varepsilon > 0$, the magnitude of a vector of T independent $\mathcal{CN}(0,1)$ random variables is contained in a shell of radius $\sqrt{T}$ and width $\varepsilon\sqrt{T}$ with probability that approaches one as $T \to \infty$ (Figure 3 displays this so-called sphere-hardening phenomenon). Hence, the ΦV and $\mathcal{CN}(0,1)$ structures for S are reconciled if $v_1 = \cdots = v_M = \sqrt{T}$ with high probability as $T \to \infty$.

4.2 Capacity lower bound

By substituting an arbitrary density for $p(S)$ in (5), one obtains a lower bound on capacity. The arguments of Section 4.1 suggest that by assigning unit probability mass to $v_1 = \cdots = v_M = \sqrt{T}$, one should obtain a lower bound, say C_l, that becomes tight as $T \to \infty$.

For $M = N = 1$,

$$\begin{aligned} C_l &= -T\log e - \log(1+\rho T) \\ &\quad - \int_0^\infty \frac{(T-1)e^{-\lambda/(1+\rho T)}\gamma\big(T-1, \frac{\rho T\lambda}{1+\rho T}\big)}{\Gamma(T)(1+\rho T)\left[\frac{\rho T}{1+\rho T}\right]^{T-1}} \log\left[\frac{(T-1)e^{-\lambda/(1+\rho T)}\gamma\big(T-1, \frac{\rho T\lambda}{1+\rho T}\big)}{(1+\rho T)\left[\frac{\rho T\lambda}{1+\rho T}\right]^{T-1}}\right] d\lambda, \end{aligned} \tag{7}$$

where $\gamma(T,z) \overset{\text{def}}{=} \int_0^z q^{T-1}e^{-q}dq$ is the *incomplete gamma* function (see [2]).

4.3 Lower bound approaches upper bound as $T \to \infty$

For $M = N = 1$ the perfect-knowledge capacity upper bound (6) is

$$C_u = T \cdot \mathrm{E}\log\big(1 + \rho|h_{11}|^2\big) = T\int_0^\infty dy\, e^{-y}\log\big(1+\rho y\big) = T(\log e)e^{1/\rho}\mathrm{E}_1(1/\rho),$$

where

$$\mathrm{E}_1(x) = \int_x^\infty \frac{e^{-y}}{y}\,dy \tag{8}$$

is the *exponential integral.* Hence, for $M = N = 1$, we expect $C/T \to (\log e)e^{1/\rho}\mathrm{E}_1(1/\rho)$ as $T \to \infty$. This intuition is made precise in the following theorem.

Theorem 3 *(Capacity, asymptotically in T) Let* $M = N = 1$. *Then* $C_l/T \leq C/T \leq C_u/T = (\log e)e^{1/\rho}\mathrm{E}_1(1/\rho)$, *and* $C_l/T \to C_u/T$ *as* $T \to \infty$.

Since we have shown that, for large T, the capacity of our communication link approaches the perfect-knowledge upper bound, we also expect the joint probability density of the elements of $S = \Phi V$ to become approximately independent $\mathcal{CN}(0,1)$ as part of the sphere-hardening phenomenon described in Section 4.1. This is the content of the next theorem.

Theorem 4 *(Distribution that achieves capacity, asymptotically in T) Let* $M = N = 1$. *For the signal that achieves capacity,* $v_1/\sqrt{T}$ *converges in distribution to a unit mass at* 1, *as* $T \to \infty$.

4.4 Capacity and capacity bounds for $M \geq 1$, $N \geq 1$, and $T = 1$

In this case, there are multiple transmitter and receiver antennas and the coherence interval is $T = 1$, corresponding to a rapidly changing channel. Theorem 1 implies that the capacity for $M \geq T$ is the same as for $M = T$, so we assume, for computational purposes, that $M = 1$. The transmitted signal is then a complex scalar having a uniformly distributed phase. All of the transmitted information is contained in the magnitude, whose probability density function was obtained with the Blahut-Arimoto algorithm.

Figure 4 displays the capacity as a function of N for arbitrary M. Since $T = 1$, the receiver cannot estimate the propagation coefficients reliably, and hence the capacity is far less than the perfect-knowledge upper bound. The difference between the capacity and the perfect-knowledge upper bound becomes especially dramatic as M and N both simultaneously increase since the upper bound grows approximately linearly with the minimum of M and N (not shown in figure), while in Figure 4 the growth of capacity with N (recall that $M \leq T = 1$) is very moderate and appears to be logarithmic.

5 Conclusions

We have taken a fresh look at the problem of communicating over a flat-fading channel using multiple-antenna arrays. No knowledge about the propagation coefficients and no ad hoc training schemes were assumed. Three key findings emerged from our research.

First, there is no point in making the number of transmitter antennas greater than the length of the coherence interval. In a very real sense, the ultimate capacity of a multiple-antenna wireless link is determined by the number of symbol periods between fades. This is somewhat disappointing since it severely limits the ultimate capacity of a rapidly fading channel. For example, in the extreme case where a fresh fade occurs every symbol period, only one transmitter antenna can be usefully employed. Strictly speaking, one could increase capacity indefinitely by employing a large number of receiver

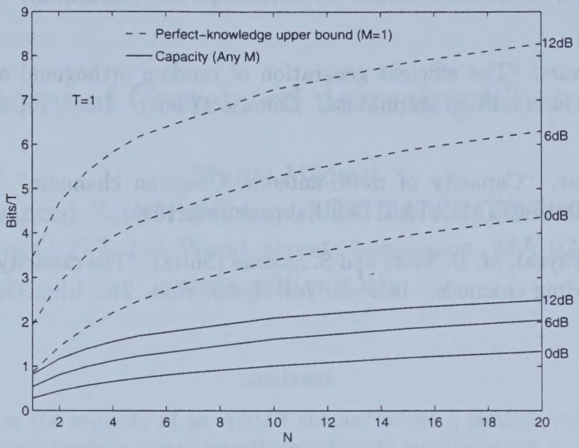


Figure 4: Capacity and perfect-knowledge upper bound versus number of receiver antennas N (SNR=0dB, 6dB, 12dB, arbitrary number of transmitter antennas, coherence interval equal to one). The gap between capacity and upper bound only widens as $N \to \infty$.

antennas, but the capacity appears to increase only logarithmically in this number—not a very effective way to boost capacity.

Second, the transmitted signals that achieve capacity are mutually orthogonal with respect to time among the transmitter antennas. The constituent orthonormal unit vectors are isotropically distributed and statistically independent of the signal magnitudes. This result provides insight for the design of efficient signaling schemes, and it greatly simplifies the task of determining capacity, since the dimensionality of the optimization problem is only equal to the number of transmitter antennas.

Third, when the coherence interval becomes large compared with the number of transmitter antennas, the normalized capacity approaches the capacity obtained as if the receiver knew the propagation coefficients. The magnitudes of the time-orthogonal signal vectors become constants that are equal for all transmitter antennas. In this regime, all of the signaling information is contained in the directions of the random orthogonal vectors, the receiver learns the propagation coefficients, and the channel becomes similar to the classical Gaussian channel.

References

[1] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[2] B. M. Hochwald and T. L. Marzetta, "Capacity of a mobile multiple-antenna communication link in a Rayleigh flat-fading environment," Tech. Rep. BL0112170-970512-09TM, Lucent Bell Laboratories, 1997.

[3] W. C. Jakes, *Microwave Mobile Communications.* Piscataway, NJ: IEEE Press, 1993.

[4] G. W. Stewart, "The efficient generation of random orthogonal matrices with an application to condition estimators," *Siam J. Numer. Anal.*, vol. 17, pp. 403–409, 1980.

[5] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," Tech. Rep. BL-0112170-950615-07TM, AT&T Bell Laboratories, 1995.

[6] I. C. Abou Faycal, M. D. Trott and S. Shamai (Shitz), "The capacity of discrete-time Rayleigh fading channels," in *Proc. Int. Symp. Info. Th.*, Ulm, Germany, 1997, p. 473.

Capacity of Correlated Jamming Channels.

Muriel Médard
Advanced Networks Group, MIT Lincoln Laboratory
Room C-277, 244 Wood Street, Lexington, MA 02173
medard@ll.mit.edu

Abstract

We derive the capacity of an AWGN channel where a jammer can make use of a (legitimate) sender's signal, intercepted through eavesdropping, to disrupt the legitimate user. Such a phenomenon is called correlated jamming. We derive the capacity of a correlated jamming channel with phase or time jitter in the channel tapping or jamming.

1 Introduction.

Correlated jamming, the situation where the jammer may correlate his signal to that of the legitimate user, poses a particular risk in certain types of channels. The subject has received attention in the wireless domain, for instance in the form of repeat-back jamming ([1, 2, 3, 4]). Correlated jamming may also yield a particularly nefarious attack in the area of optical communications, since a fiber may be physically tapped and have a signal injected into it.

Correlated jamming has also received attention in the information-theoretic literature. The effect of correlated jamming on SNR when the jammer has complete or incomplete information about the legitimate sender was examined in ([6]). Pessimum inteference was studied by Blachman ([7, 8]), who gives bounds on the capacity. A more general form of correlated jamming was studied by Başar and Başar in [9, 11, 10]. In [9], the optimum encoder-decoder structure and most nefarious noise are found unde a square-difference distortion measure for a finite number of transmissions of a Gaussian process over an AWGN channel with correlated jamming. In [11, 10], channel feedback is added. The jamming signal is subject to a power constraint but is allowed to be arbitrarily correlated with the encoded legitimate signal.

The purpose of this paper is to establish exact capacity results for channels with correlated jamming when the jammer has full or partial knowledge of the user's signal and when the jammer must contend with phase or time jitter. We use an approach combining optimal detection applied to entropy and convexity of entropy. We derive, in section 2, our results for correlated jamming, without making any restrictive assumptions about the capabilities of the jammer. The jammer is assumed to have perfect knowledge of the legitimate signal that is sent and he inserts a jamming signal by making use of his knowledge of the legitimate signal. In section 3, we consider eavesdropper correlated jamming, where the jammer derives his knowledge of the legitimate signal by eavesdropping. As we shall see, the jammer actually does not need very extensive capabilities to disrupt communications significantly. The results are congruent with the results for the Gaussian test channel studied in [9] . In sections 4 and 5, we next extend our analysis to take into account the fact that, under some conditions, coherence and synchronicity may not be maintained by the jammer. We conclude and give directions for further research in section 6.

2 The correlated jamming attack.

We first consider how the jammer can disrupt communications when it has access to the legitimate signal. Let us first establish our notation. We assume an AWGN model for the channel noise. Let T be the time interval during which we transmit. By sampling at baseband at the Nyquist rate, W, we can reduce the continuous-time channel to a discrete-time channel where WT complex symbols are sent. The input symbols form the vector $\underline{X}_{2WT} = (X[1], \cdots X[2WT])$ where $X[i]$ is the real part of the i^{th} sample of the transmitted signal, sampled at the Nyquist rate of W Hz, and $X[WT+i]$ is the imaginary part of i^{th} sample. We denote by σ_X^2 the the average energy constraint on the real (complex) component of the transmitted signal. We assume an AWGN model for the channel noise. The noise vector is $\underline{N}'_{2WT}$, where all $2WT$ symbols are real IID zero-mean Gaussian distributed with variance σ_N^2. The signal received by the legitimate user in the absence of jamming is the vector $\underline{X}_{2WT} + \underline{N}'_{WT}$.

Let the average energy constraint on the jammer over the bandwidth W and time T be E_{Jammer}. Let $\underline{J}_{2WT}$ be the signal the jammer uses to disrupt $\underline{X}_{2WT} + \underline{N}'_{2WT}$. If $\underline{J}_{2WT}$ were uncorrelated with $\underline{X}_{2WT}$, then we know that it would be selected to be AWGN ([14, pg. 337]). Given the nature of common sources of AWGN in communications channels, we impose the condition that the jamming signal $\underline{J}_{2WT}$ must be independent of the AWGN noise component $\underline{N}'_{2WT}$. The jammer inserts his jamming signal $\underline{J}_{2WT}$ into the received signal of the legitimate user. We shall denote the attenuation at the point of jamming by γ. The legitimate receiver must attempt to recover $\underline{X}_{2WT}$ from the signal

$$\underline{Z}_{2WT} = \underline{X}_{2WT} + \sqrt{\gamma}\underline{J}_{2WT} + \underline{N}'_{2WT}. \tag{1}$$

Figure 1 shows the system considered in this section. The jammer seeks to minimize

$$I\left(\underline{Z}_{2WT}; \underline{X}_{2WT}\right) = h\left(\underline{X}_{2WT}\right) - h\left(\underline{X}_{2WT} | \underline{Z}_{2WT}\right). \tag{2}$$

The jammer cannot control the first term of the RHS of (2) and must therefore strive to maximize the second term, which may be rewritten as

$$h\left(\underline{X}_{2WT} | \underline{Z}_{2WT}\right) = h\left(\underline{X}_{2WT} - A\left(\underline{Z}_{2WT}\right) | \underline{Z}_{2WT}\right) \tag{3}$$

for some constant matrix A, since a constant shift leaves entropy unchanged. Therefore, since entropy is reduced by conditioning,

$$h\left(\underline{X}_{2WT} | \underline{Z}_{2WT}\right) \leq h\left(\underline{X}_{2WT} - A\left(\underline{Z}_{2WT}\right)\right). \tag{4}$$

In particular, we may pick A to be the linear least squared error (LLSE) estimate of $\underline{X}_{2WT}$ from $\underline{Z}_{2WT}$. The variance of the estimate error $\underline{X}_{2WT} - A\left(\underline{Z}_{2WT}\right)$ is

$$\Lambda_{\underline{X}_{2WT}} - \Lambda_{(\underline{X}_{2WT}, \underline{Z}_{2WT})} \Lambda^{-1}_{\underline{Z}_{2WT}} \Lambda_{(\underline{Z}_{2WT}, \underline{X}_{2WT})}$$

which we shall denote Λ_{Error}. Since, for a given autocorrelation matrix, entropy is maximized by a Gaussian distribution, we have that ([13, pg. 234]),

$$h\left(\underline{X}_{2WT} - A\left(\underline{Z}_{2WT}\right) | \underline{Z}_{2WT}\right) \leq \frac{1}{2} ln\left((2\pi e)^{2WT} \mid \Lambda_{Error} \mid \right) \tag{5}$$

where $|\ |$ denotes the absolute value of the determinant.

Let us now establish what signalling scheme the sender should use and what jamming scheme the jammer should use. To do so, we shall first assume a WGN signal for the sender. We shall determine the optimal jamming scheme for a WGN sender's signal. The jammer attempts to minimize $I\left(\underline{Z}_{2WT}; \underline{X}_{2WT}\right)$ by changing the transition probabilities of the channel connecting $\underline{Z}_{2WT}$ and $\underline{X}_{2WT}$. Next, we shall assume that the jammer uses the jamming scheme that is optimal for a WGN sender's signal. We shall then allow the sender to vary its signal and shall determine that the maximum rate is achieved by sending a WGN signal. Thus, since the mutual information is a concave function in the input probability and a convex function in the set of transition probabilities [14, pp. 89–91], we shall have shown that the best signalling scheme for the sender is a WGN signal and shall have determined what the jammer's scheme is for that signalling case. Our approach is different from [9], which does not consider capacity, but our results are congruent with those of [9].

We first assume that the transmitter picks a Gaussian signal $\underline{X}_{2WT}$ scheme for the AWGN over which it sends. By selecting $\underline{J}_{2WT}$ to be Gaussian and jointly Gaussian with $\underline{X}_{2WT}$ and $\underline{N}'_{2WT}$, the LLSE estimate $\underline{X}_{2WT} - A\left(\underline{Z}_{2WT}\right)$ is also the minimum variance error (MVE) estimate and the Bayesian estimate. If $\underline{J}_{2WT}$ is Gaussian, then $\underline{X}_{2WT} - A\left(\underline{Z}_{2WT}\right)$ is a Gaussian vector error which is independent of the signal from which the estimate was obtained, therefore both (4) and (5) hold with equality. Moreover, any value of Λ_{Error} allowed by our constraints may be achieved by picking $\underline{J}_{2WT}$ Gaussian and jointly Gaussian with $\underline{X}_{2WT}$. Hence, the jammer will pick $\underline{J}_{2WT}$ Gaussian and jointly Gaussian with $\underline{X}_{2WT}$.

Let us now determine the form of $\Lambda_{\underline{J}_{2WT}}$ which the jammer will pick. Since conditioning reduces entropy, we may use the chain rule for entropies to write

$$h(\underline{X}_{2WT}|\underline{Z}_{2WT}) \leq \sum_{i=1}^{TW} h(X[i]|Z[i]) + \sum_{i=TW+1}^{2TW} h(X[i]|Z[i]). \tag{6}$$

The $X[i]$s are IID, the $N'[i]$s are IID and the $X[i]$s and the $N'[i]$s are mutually independent. Hence, from (1), we see that (6) may be achieved with equality by taking $J[i]$ to be independent of all $N'[j]$, $X[j]$ for all $j \neq i$. Thus, from (2)-(6), minimizing the LHS of (2) becomes equivalent to minimizing

$$\sum_{i=1}^{2WT} h\left(X[i]\right) - h\left(X[i]|Z[i]\right). \tag{7}$$

Note that $\underline{J}_{2WT}$ and $\underline{X}_{2WT}$ are Gaussian and jointly Gaussian implies that $\underline{X}_{2WT}$ and $\underline{Z}_{2WT}$ are Gaussian and jointly Gaussian. We have ([15])

$$h\left(X[i]\right) - h\left(X[i]|Z[i]\right) = \frac{1}{2}\left(ln\left(\sigma_X^2\right) + ln\left(\sigma_Z^2\right) - ln\left(\left|\begin{matrix} \sigma_X^2 & \sigma_{XZ} \\ \sigma_{XZ} & \sigma_Z^2 \end{matrix}\right|\right)\right). \tag{8}$$

Therefore, we may achieve any value of (7) by picking $J[i]$ to be the sum of $X[i]$, multiplied by some factor, and of some Gaussian variable independent of $X[i]$. To minimize the LHS of (2), we may choose for each i $J[i] = \xi_i X[i] + R[i]$. To satisfy our conditions that make (6) hold with equality, all the $R[i]$s are mutually independent and independent of all $X[j]$ for all $j \neq i$. Moreover, to satisfy the condition that the jamming signal be independent of the AWGN, we must have all the $R[i]$s independent of all $N'[j]$ for all j. Finally, since all the $X[i]$s are IID and all the $N[i]$s are IID, we may use the concavity of the ln function to establish that we want all the $R[i]$s to be IID and all the ξ_i to be equal to some ξ. We may rewrite (7) as $\sum_{i=1}^{2WT} h\left(Z[i]\right) - h\left(Z[i]|X[i]\right)$ where $Z[i] = \left(1 + \sqrt{\gamma}\xi\right) X[i] + R[i] + N'[i]$. The signal seen by the jammed user therefore has a white Gaussian signal with variance $\left(1 + \sqrt{\gamma}\xi\right)^2 \sigma_X^2$ and AWGN with variance $\sigma_R^2 + \sigma_N^2$. Thus, the problem of minimizing the LHS of (2) now becomes the problem of minimizing

$$g\left(\xi\right) = \frac{1}{2} ln \left(\frac{(1 + \sqrt{\gamma}\xi)^2 \sigma_X^2 + \gamma\sigma_R^2 + \sigma_N^2}{\gamma\sigma_R^2 + \sigma_N^2} \right) \tag{9}$$

subject to

$$\xi^2 \sigma_X^2 + \sigma_R^2 \leq E_{Jammer}. \tag{10}$$

For a problem where the jammer has very high power, where $\gamma E_{Jammer} \geq \sigma_X^2$, then we may design $\underline{X}_{2WT}$ to subtract all the signal part of $\underline{Z}_{2WT}$. If the jammer cannot altogether eliminate all of the signal component, then it must allocate its power most effectively to reduce the SNR. We have that $g''\left(\xi\right) \geq 0$ and that $g'\left(\xi\right) = 0$ only for

$$\xi = -\frac{\sigma_X^2 + \gamma E_{Jammer} + \sigma_N^2 - \mid \sigma_X^2 - \gamma E_{Jammer} - \sigma_N^2 \mid}{2\sqrt{\gamma}\sigma_X^2}. \tag{11}$$

We denote the above RHS by λ. Substituting and solving for ξ, we obtain that

$$\xi = \begin{cases} max\left(\lambda, -\sqrt{\frac{E_{Jammer}}{\sigma_X^2}}\right), & \text{if } \gamma E_{Jammer} \leq \sigma_X^2 \\ -\frac{1}{\sqrt{\gamma}}, & \text{if } \gamma E_{Jammer} \geq \sigma_X^2 \end{cases}. \tag{12}$$

Let us now show that the sender cannot do better than to send a WGN signal. Let us consider that we have an AWGN channel, with amplification $1 + \sqrt{\gamma}\xi$ and with noise variance $\gamma\sigma_R^2 + \sigma_N^2$. Then, we know that a WGN signal is optimal. Since the mutual information is concave in the input distribution and convex in the transition probabilities, we have a saddle point.

Note that the correlated jammer simply needs to be able to amplify negatively (i.e. amplify with a phase shift of π) $\underline{X}_{2WT}$ and to generate AWGN. The variable ξ denotes the factor by which the original signal $\underline{X}_{2WT}$ is multiplied to form the correlated component of the jamming signal $\underline{J}_{2WT}$. In particular, the jammer does not need to *detect* $\underline{X}_{2WT}$. The average mutual information obtained in (9) is indeed a capacity, because we have reduced the channel to an AWGN channel.

3 The eavesdropper correlated jamming attack.

We now examine the question of what signal should be sent by a jammer whose only access to the transmitted signal is via tapping. The jamming signal can never be such that observing it yields a better estimate of $\underline{X}_{2WT}$ than that obtained via tapping. The noise over the eavesdropper's channel is represented by the vector $\underline{N}_{2WT}$ where all $2WT$

symbols are real IID zero-mean Gaussian. The rate obtained via tapping by an eavesdropper can easily be obtained, for instance by modifying our problem to be compatible with the framework of [16, 17]. Note that, in the rest of the paper, we do not endow the information obtained by the eavesdropper about the legitimate transmitted signal with any semantic implications. A correlated jammer does not need to have information about the plaintext, but only about the signal sent by the legitimate user.

Figure 2 shows the eavesdropper correlated channel considered in this section. The attenuation at the point of jamming is denoted by γ and the attenuation at the point of tapping by α. The jammer sends a signal J. Using our previous convention for representing a complex signal, we may express the i^{th} sample of J as:

$$J[i] = \xi Y[i] + R'[i]. \tag{13}$$

The channel jammed by an eavesdropper correlated jammer sees the signal whose i^{th} time sample is given by

$$Z[i] = \sqrt{\alpha}J[i] + N'[i] = (1 + \sqrt{\gamma}\sqrt{\alpha}\xi)X[i] + \sqrt{\alpha}\xi N[i] + \sqrt{\alpha}R'[i] + N'[i]. \tag{14}$$

For ease of exposition, we assume that the variance of $N'[i]$ and that of $N[i]$ are both equal to σ_N^2. The extension to the case where they are different is straightforward. The same arguments as in the previous section show that a saddle point is reached when the sender and the jammer transmit in such a manner that the $X[i]$s are samples of a WGN process and the $R'[i]$s are also samples of a WGN process and are independent of the X and N processes. Thus, the RHS of (14) has a signal term with a white Gaussian distribution of variance $(1+\sqrt{\gamma}\sqrt{\alpha}\xi)^2\sigma_X^2$ and an AWGN component with variance $\gamma\sigma_{R'}^2 + (\gamma\alpha\xi^2+1)\sigma_N^2$. The coefficient ξ now indicates the weighting that the jammer gives to the signal $\underline{Y}_{2WT}$ obtained through tapping. Therefore, the function to minimize, which is given by (9) in the previous section, is now

$$g_{eaves}(\xi) = \frac{1}{2}ln\left(\frac{(1+\sqrt{\gamma}\sqrt{\alpha}\xi)^2\sigma_X^2 + \gamma\sigma_{R'}^2 + (\gamma\alpha\xi^2+1)\sigma_N^2}{\gamma\sigma_{R'}^2 + (\gamma\alpha\xi^2+1)\sigma_N^2}\right). \tag{15}$$

The constraint of (10) becomes

$$\xi^2\alpha\left(\sigma_X^2 + \sigma_N^2\right) + \sigma_{R'}^2 \leq E_{Jammer}. \tag{16}$$

We may verify that any solution feasible in the eavesdropper correlated case is feasible in the correlated case studied in section 2. The minimization of (9) subject (10) cannot therefore be more than the solution to minimizing (15) subject to (16) alone. The solution of (15) subject to (16) is

$$\xi = \begin{cases} -\frac{1}{\sqrt{\gamma\alpha}}, & \text{if } \gamma E_{Jammer}\frac{\sigma_X^2}{\sigma_X^2+\sigma_N^2} \geq \sigma_X^2 \\ \mu, & \text{if } \gamma E_{Jammer}\frac{\sigma_X^2}{\sigma_X^2+\sigma_N^2} \geq \sigma_X^2, \text{ and } g_{eaves}(\mu) < g_{eaves}(\nu) \\ \nu & \text{otherwise} \end{cases} \tag{17}$$

for $\nu = -\sqrt{\frac{E_{Jammer}}{\alpha(\sigma_X^2+\sigma_N^2)}}$ and $\mu = \frac{-b+\sqrt{b^2-4ac}}{2a}$ where $a = \sigma_X^2\gamma\alpha$, $b = \sqrt{\gamma\alpha}\sigma_N^2 + \sqrt{\alpha}\gamma^{\frac{3}{2}}E_{Jammer} + \sqrt{\gamma\alpha}\sigma_X^2$ and $c = \sigma_N^2 + \gamma E_{Jammer}$. As for the correlated jamming case studied in the previous section, we see that the capabilities required for performing eavesdropper correlated jamming are amplification and phase shift of the tapped signal and addition of white noise. The conditions we impose on the jammer in (16) could be changed to take into account different constraints. For instance, a minimum noise constraint of σ_{min}^2 for the jamming signal would lead to a constraint of the form $\sigma_{R'}^2 \geq \sigma_{min}^2$. A maximum amplification amplitude of ξ_{max} would lead to a constraint of the form $\xi^2 \leq \xi_{max}^2$.

4 The effect of phase jitter.

In this section, we analyze the effect upon correlated jamming of a phase shift which is not known to the jammer. If the phase shift were known to the jammer, he could compensate for it in his jamming scheme. Such a phase shift may occur because the jammer has mistakingly evaluated the phase of the legitimate user or it may be deliberately introduced by the legitimate user after the point of tapping. Such a phase may be modelled by adding a phase shift to the legitimate signal after the eavesdropping has occured or by adding a phase shift to the jamming signal. We choose the latter approach. We denote the phase shift by ϕ. We assume that the phase shift is time-invariant, therefore we do not have bandwidth expansion which is due to changes in the phase shift. Moreover, the fact that the phase shift is constant implies that the legitimate receiver is informed of the phase shift or can determine it without affecting capacity. Therefore, the receiver can take ϕ into account when he receives the signal.

We first look at the effect of a phase shift for a given ξ and then assume a distribution fo ξ. The received signal sampled vector, $\underline{Z}_{2WT}$, is given by

$$\begin{aligned} Z[i] &= X[i] + \sqrt{\gamma}\sqrt{\alpha}\xi\left(cos\left(\phi\right)X[i] - sin\left(\phi\right)X[i+WT]\right) \\ &+ \sqrt{\gamma}\sqrt{\alpha}\xi N[i] + \sqrt{\gamma}R'[i] + N'[i] \end{aligned} \tag{18}$$

for $1 \leq i \leq WT$ and

$$\begin{aligned} Z[i] &= X[i] + \sqrt{\gamma}\sqrt{\alpha}\xi\left(cos\left(\phi\right)X[i] + sin\left(\phi\right)X[i-WT]\right) \\ &+ \sqrt{\gamma}\sqrt{\alpha}\xi N[i] + \sqrt{\gamma}R'[i] + N'[i] \end{aligned} \tag{19}$$

for $WT+1 \leq i \leq 2WT$. We may compute the capacity of the legitimate user for a given ϕ and ξ. As in our previous sections, we first assume that the sender's signal is complex WGN. The same arguments as in the previous sections lead to the fact that the jammer chooses R' to be AWGN uncorrelated with N and X. The entropy of $\underline{Z}_{2WT}$ is $\frac{1}{2}ln\left((2\pi)^{2WT} \mid \Lambda_{\underline{Z}_{2WT}} \mid\right)$, where $\Lambda_{\underline{Z}_{2WT}}$ is a diagonal matrix with diagonal terms

$$d = \sigma_X^2\left(1 + 2\sqrt{\gamma}\sqrt{\alpha}\xi cos\left(\phi\right) + \gamma\alpha\xi^2\right) + \left(\gamma\alpha\xi^2 + 1\right)\sigma_N^2 + \gamma\sigma_{R'}^2. \tag{20}$$

Therefore, from (18-19), the mututal information for the channel seen by the legitimate user is given by

$$WTln\left(\frac{\mid d \mid}{\left(1+\gamma\alpha\xi^2\right)\sigma_N^2 + \gamma\sigma_{R'}^2}\right). \tag{21}$$

Note that, as in the previous sections, the optimum signalling for the user when the jammer uses such a scheme is WGN. We have therefore found a saddle point.

If the jammer knows that there is phase shift and knows a distribution on the phase shift, he may take the phase shift into account in his choice of ξ. The goal of the jammer is to minimize the expected capacity of the legitimate user's channel, hence to minimize $E_\phi\left[I\left(\underline{Z}_{2WT}; \underline{X}_{2WT}|\phi\right)\right]$, where the expectation is taken over ϕ, which is known to the receiver. We now examine the case where the *a priori* distribution of ϕ is uniform between 0 and 2π, i.e. all guesses about the phase by the jammer are equally poor. Using (20, 21), the goal of the jammer is to minimize

$$\frac{1}{4\pi}\int_0^{2\pi} ln\left(1 + \frac{\sigma_X^2\left(1 + 2\sqrt{\gamma\alpha}\xi cos\left(\phi\right) + \gamma\alpha\xi^2\right)}{\left(1+\gamma\alpha\xi^2\right)\sigma_N^2 + \gamma\sigma_{R'}^2}\right) d\phi. \tag{22}$$

By a change of variables, we use equation 4.292.3 in [20] to rewrite (22) as

$$g_{phase_1}(\xi) = \frac{1}{2} ln \left(a \frac{1 + \sqrt{1 - \frac{b^2}{a^2}}}{2} \right) \tag{23}$$

where $a = 1 + \frac{\sigma_X^2(1+\gamma\alpha\xi^2)}{(1+\gamma\alpha\xi^2)\sigma_N^2+\gamma\sigma_{R'}^2}$ and $b = \frac{2\sigma_X^2\sqrt{\gamma\alpha}\xi}{(1+\gamma\alpha\xi^2)\sigma_N^2+\gamma\sigma_{R'}^2}$. We have $g'_{phase_1}(\xi) = 0$ and $g''_{phase_1}(\xi) > 0$ only for $\xi = 0$, therefore the jammer chooses ξ to be 0. Thus, uncorrelated jamming is preferable when the jammer has abolutely no estimate of ϕ.

5 The effect of timing jitter.

Instead of a phase offset, there may exist a timing offset introduced by the legitimate user or caused by the jammer's inability to time the jamming signal exactly. For instance, the fiber lengths may be such that a jammer cannot exactly achieve synchronization, because the jamming signal traverses a longer length of fiber than the legitimate signal. Note that, in general, a timing jitter and a phase offset will occur together. We do not consider the two together in this paper, but the techniques we develop would be applicable. Even if the jammer knows the offset exactly, he cannot compensate for such a time offset because it is not reversible, unlike a phase offset. If the legitimate user is delayed and the jammer knows this delay exactly, the jammer can introduce the same delay in the jamming signal. Let us suppose that the jammer is lagging behind the legitimate signal, perhaps because the jammer uses a piece of fiber which is slightly longer than that traversed by the legitimate signal. The received signal sampled vector, $\underline{Z}_{2WT}$, is given by

$$Z[i] = X[i] + \sqrt{\gamma}\sqrt{\alpha}\xi\left(\beta X[i] + (1-\beta)X[i-1]\right) + \sqrt{\gamma}\sqrt{\alpha}\xi N[i] + \sqrt{\gamma}R'[i] + N'[i] \tag{24}$$

for $1 \leq i \leq 2WT$. Note that the timing offset would involve $X[i+1]$ instead of $X[i]$ if the legitimate user introduced a delay after the tapping point.

As before, we assume that the legitimate user sends a Gaussian signal, although the signal is no longer white, since we have an ISI channel. Given such a signalling scheme, the same arguments used in the previous sections imply that the best jammer policy is to send a jamming signal which is the sum of the negatively amplified tapped signal and of AWGN uncorrelated with X and N'. Given such a jamming policy, we may show that a Gaussian signal is the best that the sender can use. In order to find the optimal choice of ξ, we shall derive the capacity of the channel in terms of ξ. The terms of the autocorrelation matrix of the received signal are

$$\Lambda_{\underline{Z}_{2WT}}[i,j] = \begin{cases} p, & \text{for } i = j \\ q, & \text{for } \mid i - j \mid = 1 \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

$$\begin{aligned} p &= \sigma_X^2\left(\left(1+\sqrt{\gamma}\sqrt{\alpha}\beta\xi\right)^2 + (1-\beta)^2\right) + \sigma_N^2\left(\gamma\alpha\xi^2+1\right) + \gamma\sigma_{R'}^2 \\ q &= \sigma_X^2\left(1+\sqrt{\gamma}\sqrt{\alpha}\beta\xi\right)(1-\beta). \end{aligned} \tag{26}$$

If ξ and β are given, then the capacity of the channel is known, because it is a channel with time-invariant ISI ([18, 19, 5]), which can equivalently be considered as a channel with colored Gaussian noise ([13]). We shall use the results of ([18]) to find the capacity of the correlated jamming channel. Let us define the spectrum of the Toeplitz matrix $\Lambda_{\underline{Z}_{2WT}}$ to be

$$| a(\lambda) |^2 = \left| \left(1 + \sqrt{\gamma}\sqrt{\alpha}\xi\beta\right) + \sqrt{\gamma}\sqrt{\alpha}\xi\,(1-\beta)\, e^{i\lambda} \right|^2 . \tag{27}$$

Let us define $\sigma^2 = \sigma_N^2 \left(\gamma\alpha\xi^2 + 1\right) + \gamma\sigma_{R'}^2$. From (5.3) in ([18]), we have that the capacity is given by

$$\frac{1}{4\pi}\int_0^{2\pi} ln\left(\max\left[| a(\lambda) |^2 \theta\sigma^{-2}, 1\right]\right) d\lambda \tag{28}$$

where $\sigma_X^2 = \frac{1}{2\pi}\int_0^{2\pi} \max\left[\theta - \sigma^2 \, | a(\lambda) |^{-2}, 0\right] d\lambda$. In order to obtain a closed form solution, let us assume that the SNR is large enough that

$$\theta = \sigma_X^2 + \frac{\sigma^2}{(1+c)^2}\frac{1}{2\pi}\int_0^{2\pi} \frac{1}{1 + 2\frac{b}{1+c} cos\,(\lambda) + \left(\frac{b}{1+c}\right)^2} d\lambda \tag{29}$$

for $b = \sqrt{\gamma\alpha}\xi\beta$ and $c = \sqrt{\gamma\alpha}\xi\,(1-\beta)$. If we assume that $\left| \frac{b}{1+c} \right| \leq 1$, we may use, from [20], equation 3.665.2. to obtain

$$g_{delay}\,(\xi) = \frac{1}{2}\left[ln\left(\theta\sigma^{-2}\right) + ln\left((1+c)^2\right)\right]. \tag{30}$$

Note that our average mutual information does indeed give us a capacity, because we have reduced the channel to a time-invariant ISI chanel with AGWN.

6 Conclusions.

We have investigated the capacity of channels with correlated jamming. We have shown that correlated jamming can be a particularly nefarious attack, when compared to uncorrelated jamming. Correlated jamming can occur even when the jammer has access to only a noisy version of the legitimate user's signal and when the jammer must contend with phase or timing jitter. The implications of in-fiber correlated jamming for optical communications may be significant because, as we have shown, correlated jamming requires very simple processing on the part of the jammer and may be effective without changing the output power that would be received for an uncorrupted legitimate signal. Therefore, most traditional methods for detecting fiber tapping and jamming, which rely upon detecting variations in received power, may not be effective.

References

[1] E.A. Gerianotis, "Direct-Sequence Spread-Spectrum Communications in a Multiple-Tone and Repeat-Back Jamming Environment", in *Proceedings of the 1983 IEEE Military Communications Conference*, vol. 3, pp. 765–769.

[2] F. Qian, B.D. Van Veen, "Partially Adaptive Beamforming for Correlated Interference Rejection", *IEEE Transactions on Signal Processing*, vol. 43, no. 2, February 1995, pp. 506–515.

[3] L.C. Godara, "Beamforming in the Presence of Broadband Correlated Arrivals", *Journal of the Acoustical Journal of America*, vol. 92, no. 5, November 1992, pp .2702-2708.

[4] P. Evans, *Measured Effects of Repeater Jamming on Direct-Sequence Spread-Spectrum Receivers that Use Envelope Detectors*, M.S. Thesis, Naval Postgraduate School, Monterey, California, September 1989.

[5] W. Hirt, J.L. Massey, "Capacity of the Discrete-Time Gaussian Channel with Intersymbol Interference", *IEEE Transactions on Information Theory*, vol. 34, no. 3, May 1988, pp. 380–388.

[6] L.-H. Zetterberg, "Signal Detection Under Noise Interference in a Game Situation", *IRE Transactions on Information Theory*, vol. IT-8, no. 1, September 1962, pp. S47–S51.

[7] N.M. Blachman, "On the Capacity of a Band-Limited Channel Perturbed by Statistically Dependent Interference", *IRE Transactions on Information Theory*, vol. IT-8, no. 1, January 1962, pp. 48–55.

[8] N.M. Blachman, "The Effect of Statistically Dependent Interference Upon Channel Capacity", *IRE Transactions on Information Theory*, vol. IT-8, no. 1, September 1962, pp. S53–S57.

[9] T. Başar, "The Gaussian Test Channel with an Intelligent Jammer", *IEEE Transactions on Information Theory*, vol. IT-29, no. 1, January 1983, pp. 152–157.

[10] T. Ü. Başar , T. Başar, "Optimum Linear Causal Coding Schemes for Stochastic Processes in the Presence of Correlated Jamming", in *Proceedings of the Twentieth Conference on Information Sciences and Systems*, 1986, pp. 384–386.

[11] T. Başar, T. Ü. Başar "A Bandwidth Expanding Scheme for Communication Channels with Noiseless Feedback in the Pesence of Unknown Jamming Noise", *Journal of the Franklin Institute*, vol. 317, no. 2, pp. 73–78.

[12] T. Ü. Başar , T. Başar, "Robust Linear Coding in Continuous-time Communication Systems in the Presence of Jamming and with Noisy Side Information at the Decoder", in *Proceedings of the 1982 Conference on Information Science Systems*, 1982, pp. 323–326.

[13] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[14] R.G. Gallager, *Information Theory and Reliable Communications*, John Wiley & Sons, 1968.

[15] M.S. Pinsker, "Information About a Gaussian Random Stationay Process Contained in a Another Process, Stationarily Correlated With the First", Doklady Akademii Nauk, tome XCIX, no. 2, 1954, pp. 213–216.

[16] A.D. Wyner, "The Wire-Tap Channel", *The Bell System Technical Journal*, vol. 54, no. 8, October 1975, pp. 1355-1387.

[17] L.H. Ozarow, A.D. Wyner, "Wire-Tap Channel II", in *Proceedings of Eurocrypt 84*, Springer-Verlag.

[18] R.M. Gray, "On the Asymptotic Eigenvalue Distribution of Toeplitz Matrices", *IEEE Transactions on Information Theory*, vol. IT-18, no. 6, November 1972, pp. 725–729.

[19] B.S. Tsybakov, "On the Transmission Capacity of a Discrete-Time Gaussian Channel with Filter", Problemy Peredachi Informatzii, vol. 6, 1970, pp. 78–82.

[20] I.S. Gradshteyn, I.M. Ryzhik, "Table of Integrals, Series and Products", Academic Press, 1965.

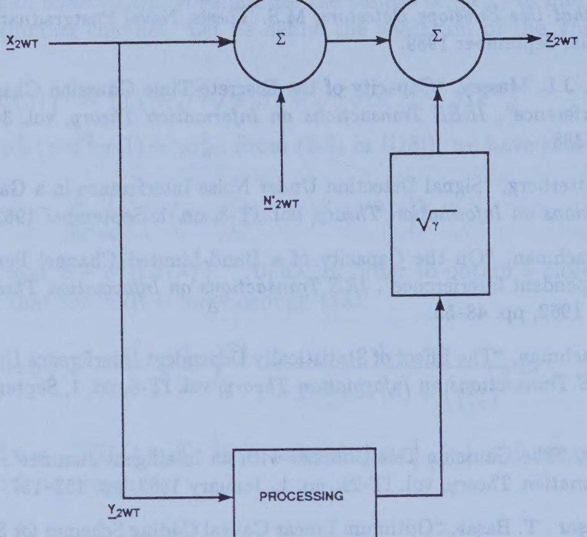

X2WT
Σ
Σ
Z2WT
N'2WT
√γ
PROCESSING
Y2WT

Figure 1: Correlated jamming channel.

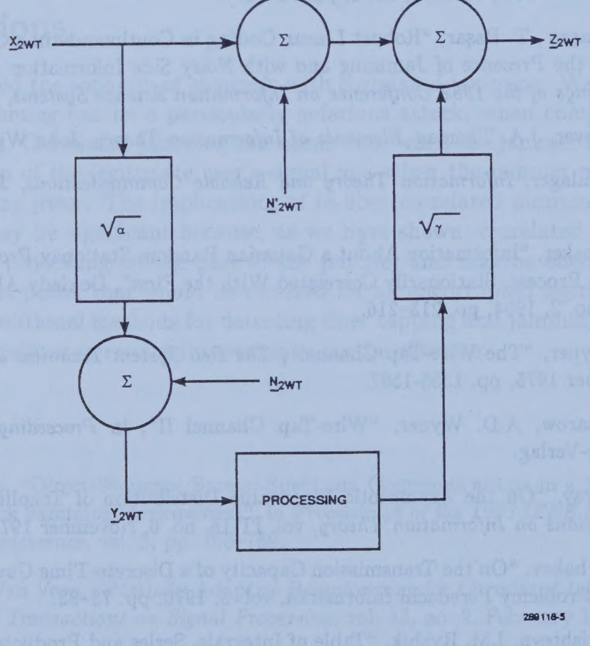

X2WT
Σ
Σ
Z2WT
√α
N'2WT
√γ
Σ
N2WT
PROCESSING
Y2WT
289118-5

Figure 2: Eavesdropper correlated jamming channel.

VOLTERRA FEEDBACK SYNTHESIS: A SYSTEMATIC ALGORITHM FOR SIMPLIFIED VOLTERRA CONTROLLER DESIGN AND REALIZATION

Patrick M. Sain
Hughes Aircraft Company
Building E1, Mailstop D115
2000 E. El Segundo Blvd.
El Segundo, CA 90245
sain@tcville.es.hac.com

Michael K. Sain and B. F. Spencer, Jr.
Structural Dynamics and Control/
Earthquake Engineering Laboratory
275 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556
sain@krizevac.ee.nd.edu
spencer@euler.ce.nd.edu

Abstract

Volterra feedback synthesis (VFS) is an algorithm that uses a finite consecutive number of the Volterra kernels of the plant and the desired closed-loop input-output map to incorporate higher-order information about the plant into a systematic nonlinear controller design and realization. A striking feature of this method is that it permits the specification of a nonlinear desired closed-loop behavior. The control design takes place in the frequency domain and is realized as an interconnected set of linear systems. By exploiting the inherent recursive structures present in VFS controllers, the number of linear systems required for realization and the complexity of their interconnections is significantly reduced, greatly simplifying the controller design and structure. The simplified form lends itself to an automated design and simulation procedure, something that was not possible before. The procedure is demonstrated on a regulator design for an single degree of freedom structure with hysteretic damping.

1 Introduction

In 1986, Al-Baiyat and Sain [1, 2] introduced the use of Volterra operators for nonlinear regulator design in the context of the Total Synthesis Problem (TSP). Since then, Sain, *et al.*, [3, 4] utilized the TSP framework to apply Volterra operators in nonlinear servomechanism design, and in 1995, Doyle, *et al.*, [5] cast the method into a model predictive control design. Al-Baiyat [6] showed the Volterra operators comprising his controller can be realized as interconnections of linear systems. However realizations obtained via a straightforward application of Al-Baiyat's approach are unnecessarily large, in the sense that many states possess identical dynamic behavior.

This paper presents a simplified form for Al-Baiyat's realizations, by removing redundant states. The new form is equivalent to the original, and possesses a systematic and predictable structure.

First, an overview of TSP is presented, followed by a summary of relevant design and realization results for Volterra controllers from the work of Al-Baiyat and Sain. Next, the realization of the partial linearization closed-loop design proposed by Al-Baiyat is given, wherein a finite consecutive number of the Volterra kernels of the closed loop are specified to be zero, beginning with the second order kernel. Al-Baiyat's realization for the partial linearization controller design is shown to possess a simplified and predictable form that permits the controller design for arbitrarily high order to be carried out in an automatic manner.

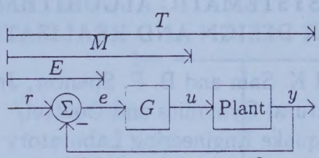


Figure 1: TSP regulator configuration.

$$\begin{tikzcd} & R \arrow[dl, dashed, "M"'] \arrow[dr, dashed, "T"] & \\ U \arrow[rr, "P"] & & Y \end{tikzcd}$$

Figure 2: Commutative design diagram.

2 Problem Definition Using Total Synthesis

The frequency domain Volterra controller design proposed by Al-Baiyat and Sain is set within the framework of the TSP paradigm described by Peczkowski and Sain [7–9]. Because the contribution of this paper centers on deriving and simplifying systematic and predictable expressions of such controllers, space constraints permit only a cursory description of TSP and the derivation of the Volterra controller design equations, so that the main points of this work can be presented in sufficient detail. In particular, consideration of TSP regulator design is sufficient; the extension of these results to TSP servomechanism design is straightforward [9].

Let R, U and Y denote real vector spaces of dimensions p, m and p, respectively, where R is the space of requests, U is the space of plant inputs, and Y is the space of plant outputs. Let $P : U \to Y$ denote an input-output description of a nonlinear plant. Define the desired closed-loop response to a command by $T : R \to Y$, and the desired plant input for a command by $M : R \to U$. The operators P, T, M and E are assumed to have Volterra series representations at the point or in the region of operation. An illustration of these operators for a regulator design is given in Fig. (1). In general, a pair (M, T) is desired such that $T = P \circ M$ and the diagram in Fig. (2) commutes. In the sequel, T is given, and the objective is to find and realize a controller $G : Y \to U$ such that $M = G \circ E$, where $E : R \to Y$, and in this case, Y is the space of output errors.

3 Plant Representation

Within the TSP framework, consider linear analytic plants, that is, plants for which the state and output equations are given by $\dot{x} = f(x) + g(x)u$, $y = h(x)$, $x(0) = x_0$, where $x \in X$, $u \in U$, $y \in Y$, $f, g : X \to X$, $g(x) : U \to X$, $h : X \to Y$, and f, g and h are analytic in x. Such plants can be represented using a so-called bilinear approximation, obtained by denoting the approximations of the state and output equations $f_1(x, u)$, $f_2(x)$, respectively, and taking a multivariable Taylor series expansion about the operating point (x_0, u_0), truncating all terms of order higher than n, yielding

$$f_1(x_0, u_0) = \sum_{i=1}^{n} A_{1i}(x_0, u_0)\tilde{x}^{[i]} + \sum_{i=1}^{n-1} D_{1i}(x_0, u_0)\tilde{x}^{[i]} \otimes \tilde{u} + D_{10}\tilde{u}, \quad f_2(x_0) = \sum_{i=1}^{n} C_{1i}(x_0, u_0)\tilde{x}^{[i]},$$

where $\tilde{x} = x - x_0$, $\tilde{u} = u - u_0$ and $x^{[i]} = x \otimes \cdots \otimes x$, the i-fold Kronecker tensor product of x with itself. To reduce notational complexity, $\otimes$ is assumed to have precedence over matrix and scalar multiplication. Suppressing the explicit dependence upon (x_0, u_0) and if f_1^j denotes $\partial^j f_1 / \partial x^j$, then $f_1^j = \sum_{i=j}^{n} A_{ji}\tilde{x}^{[i]} + \sum_{i=j}^{n-1} D_{j(i-1)}\tilde{x}^{[i]} \otimes \tilde{u}$, or

$$\frac{d}{dt}\begin{bmatrix} \tilde{x} \\ \tilde{x}^{[2]} \\ \vdots \\ \tilde{x}^{[n]} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ 0 & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{x}^{[2]} \\ \vdots \\ \tilde{x}^{[n]} \end{bmatrix} + \begin{bmatrix} D_{11} & D_{12} & \cdots & D_{1(n-1)} \\ D_{21} & D_{22} & \cdots & D_{2(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{n(n-1)} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{x}^{[2]} \\ \vdots \\ \tilde{x}^{[n]} \end{bmatrix} \otimes \tilde{u} + \begin{bmatrix} D_{10} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tilde{u},$$

for $2 \leq j \leq n$, or, more compactly, $\dot{\tilde{\bar{x}}} = A\tilde{\bar{x}} + D\tilde{\bar{x}} \otimes \tilde{u} + B\tilde{u}$, $y = C\tilde{\bar{x}}$, $\tilde{\bar{x}}(0) = \tilde{\bar{x}}_0$. In the sequel, the notation A_k denotes the matrix partition comprised of the leftmost k partitions of the top k rows of A; C_k denotes the leftmost k partitions of C; and D_k denotes the leftmost $k-1$ partitions in the top k rows of D, with $D_1 = D_{10}$.

For a multiple-input, multiple-output, finite-dimensional, causal, time-invariant plant, define the homogenous multilinear Volterra operator

$$P_i[u(t)] = \int_0^t \cdots \int_0^{\tau_i - 1} p_i(\tau_1, \ldots, \tau_i) u(t-\tau_1) \otimes \cdots \otimes u(t-\tau_i) d\tau_i \ldots d\tau_1,$$

where $t \geq \tau_1 \geq \cdots \geq \tau_i \geq 0$, and p_i is its ith Volterra kernel. A Volterra representation then has the form [1,2] $y(t) = \sum_{i=1}^{\infty} P_i[u(t)]$. The convolutional nature of this representation makes it expedient to work in the transform domain. Define the multidimensional Laplace transform of the ith Volterra kernel as [10]

$$P_i(s_1, \ldots, s_i) = \int_0^{\infty} \cdots \int_0^{\infty} p_i(\tau_1, \ldots, \tau_i)\, \mathrm{e}^{-(s_1\tau_1 + \cdots + s_i\tau_i)}\, d\tau_1 \ldots d\tau_i.$$

Al-Baiyat and Sain [9] showed that for a stationary time invariant bilinear system of the form given above, where $\overline{P}_i^j = (s_1 + \cdots + s_j)I - A_i)^{-1} D_i$, the transforms of the first three kernels and the general form are $P_1(s) = C_1\overline{P}_1^1$, $P_2(s_1, s_2) = C_2\overline{P}_2^2\{\overline{P}_1^1 \otimes I_m\}$,

$$P_3(s_1, s_2, s_3) = C_3\overline{P}_3^3\{[\overline{P}_2^2\{\overline{P}_1^1 \otimes I_m\}] \otimes I_m\}, \tag{1}$$

$$P_j(s_1, \ldots, s_j) = C_j\overline{P}_j^j\{[\overline{P}_{j-1}^{j-1}\{[\cdots \overline{P}_2^2\{\overline{P}_1^1 \otimes I_m\} \cdots] \otimes I_m\}] \otimes I_m\}. \tag{2}$$

4 Controller Design

Now, the goal is to formulate an expression for the Volterra kernels of the controller G in the frequency domain in terms of the kernels of the plant, given above, and those of the desired closed loop map, T. The development follows that given in [9]. Similarly to the Volterra operator $P_i[u(t)]$, define Volterra operators associated with the maps T and M such that $y(t) = \sum_{j=1}^{\infty} T_j[r(t)]$ and $u(t) = \sum_{k=1}^{\infty} M_k[r(t)]$. To obtain a relation between the operators P_i, T_j and M_k, replace the request signal $r(t)$ by $cr(t)$, where c is an arbitrary constant, and since Volterra operators are multilinear, equating expressions for y yields $\sum_{j=1}^{\infty} c^j T_j[r(t)] = \sum_{i=1}^{\infty} P_i(\sum_{j=1}^{\infty} c^j M_j[r(t)])$. Let r_i denote $r(t-\tau_i)$ and define

$$\sum_{i=1}^{\infty} P_i\left(\sum_{j=1}^{\infty} c^j M_j[r(t)]\right) = \sum_{j_1=1}^{\infty} \cdots \sum_{j_i=1}^{\infty} c^{j_1+\cdots+j_i} P_i\big(M_{j_1}[r(t)], \ldots, M_{j_i}[r(t)]\big), \tag{3}$$

where

$$P_i\big(M_{j_1}[r(t)], \ldots, M_{j_i}[r(t)]\big) = \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_i - 1} p_i(\tau_1, \ldots, \tau_i) M_{j_1}(r_1) \otimes \cdots \otimes M_{j_i}(r_i)\, d\tau_i \ldots d\tau_1.$$

Equating output equation expressions and suppressing the argument $[r(t)]$ yields

$$\sum_{j=1}^{\infty} c^j T_j = \sum_{i=1}^{\infty} \left(\sum_{j_1=1}^{\infty} \cdots \sum_{j_i=1}^{\infty} c^{j_1+\cdots+j_i} P_i(M_{j_1}, \ldots, M_{j_i})\right). \tag{4}$$

Equating powers of c on both sides of Eq. (4) gives $T_1 = P_1M_1$, $T_2 = P_1M_2 + P_2(M_1, M_1)$, $T_3 = P_1M_3 + P_2(M_1, M_2) + P_2(M_2, M_1) + P_3(M_1, M_1, M_1)$, and for $\kappa_i = \sum_{j=1}^{i} k_j$, $\kappa_0 = 0$,

$$T_i = P_1M_i + \sum_{j=2}^{i} \left(\sum_{k_1=1}^{i-j+1} \sum_{k_2=1}^{i-j-k_1+2} \cdots \sum_{k_{j-1}=1}^{i-\kappa_{j-2}-1} P_j(M_{k_1}, M_{k_2}, \ldots, M_{k_{j-1}}, M_{i-\kappa_j-1})\right).$$

Applying the multidimensional Laplace transform yields $T_1(s) = P_1(s)M_1(s)$, $T_2(s_1, s_2) = P_1(s_1 + s_2)M_2(s_1, s_2) + P_2(s_1, s_2)[M_1(s_1) \otimes M(s_2)]$,

$$\begin{aligned} T_3(s_1, s_2, s_3) &= P_1(s_1 + s_2 + s_3)M_3(s_1, s_2, s_3) + P_2(s_1, s_2 + s_3)\big[M_1(s_1) \otimes M_2(s_2, s_3)\big] \\ &+ P_2(s_1 + s_2, s_3)\big[M_2(s_1, s_2) \otimes M_1(s_3)\big] + P_3(s_1, s_2, s_3)\big[M_1(s_1) \otimes M_1(s_2) \otimes M_1(s_3)\big], \end{aligned} \tag{5}$$

$$\begin{aligned} T_i(s_1, \ldots, s_i) &= P_1(s_1 + \cdots + s_i)M_i(s_1, \ldots, s_i) \\ &+ \sum_{j=2}^{i} \Bigg\{ \sum_{k_1=1}^{i-j+1} \; \sum_{k_2=1}^{i-j-k_1+2} \cdots \sum_{k_{j-1}=1}^{i-\kappa_{j-2}-1} P_j(s_1 + \cdots + s_{\kappa_1}, s_{\kappa_1+1} + \cdots + s_{\kappa_2}, \ldots, s_{\kappa_{j-1}+1} + \cdots + s_i) \\ &\quad \big[M_{k_1}(s_1, \ldots, s_{\kappa_1}) \otimes M_{k_2}(s_{\kappa_1+1}, \ldots, s_{\kappa_2}) \otimes \cdots \otimes M_{i-\kappa_{j-1}}(s_{\kappa_{j-1}+1}, \ldots, s_i)\big] \Bigg\}. \end{aligned} \tag{6}$$

Observe that if the pair $(M_1(s), T_1(s))$ is chosen for $P_1(s)$, then one can proceed to the design equation involving $P_2(s_1, s_2)$, where the pair $(M_2(s_1, s_2), T_2(s_1, s_2))$ is chosen, and so on. To complete the design, consider the synthesis of the controller G that realizes the desired mappings (M, T) shown in Fig. (1). Note that the similarity of the relation $M = G \circ E$ permits expressions for the operators M_j to be obtained in a manner similar to the one for the operators T_i [9]. Also from Fig. (1), note that $e = r - y$, which can be written as $\sum_{i=1}^{\infty} E_i[r(t)] = I[r(t)] - \sum_{i=1}^{\infty} T_i[r(t)]$, where I is the identity operator, and thus $E_1(s) = I - T_1(s)$, and $E_i(s_1, \ldots, s_i) = -T_i(s_1, \ldots, s_i)$, for $i > 1$. A direct relation between the Volterra kernels of the operators T, M and G can be found if the design is well-posed on the first order, or linear, level, implying that $E_1(s)$ has an inverse. As T is usually open to choice, the latter requirement is reasonable.

5 Simplified Partial Linearization Controller Design

Consider designing a feedback system for a given nonlinear system so that $T_i = 0$, $1 < i \leq n$, where T_i is the ith Volterra kernel of the closed loop. Such a design is called *partial linearization*, and it has the desireable attribute of greatly lessening the complexity of the controller design equations.

Suppose for T that the second and higher order kernels are zero. Then, if $P_1(s)$ inverts, $M_1(s) = P_1^{-1}(s)T_1(s)$, $M_2(s_1, s_2) = -P_1^{-1}(s_1 + s_2)P_2(s_1, s_2)\big[M_1(s_1) \otimes M_1(s_2)\big]$,

$$\begin{aligned} M_3(s_1, s_2, s_3) &= -P_1^{-1}(s_1 + s_2 + s_3)\big\{P_2(s_1, s_2 + s_3)\big[M_1(s_1) \otimes M_2(s_2, s_3)\big] \\ &+ P_2(s_1 + s_2, s_3)\big[M_2(s_1, s_2) \otimes M_1(s_3)\big] + P_3(s_1, s_2, s_3)\big[M_1(s_1) \otimes M_1(s_2) \otimes M_1(s_3)\big]\big\}, \end{aligned} \tag{7}$$

$$\begin{aligned} M_i(s_1, \ldots, s_i) &= -P_1^{-1}(s_1 + \cdots + s_i) \\ &\sum_{j=2}^{i} \Bigg\{ \sum_{k_1=1}^{i-j+1} \; \sum_{k_2=1}^{i-j-k_1+2} \cdots \sum_{k_{j-1}=1}^{i-\kappa_{j-2}-1} P_j(s_1 + \cdots + s_{\kappa_1}, s_{\kappa_1+1} + \cdots + s_{\kappa_2}, \ldots, s_{\kappa_{j-1}+1} + \cdots + s_i) \\ &\quad \big[M_{k_1}(s_1, \ldots, s_{\kappa_1}) \otimes M_{k_2}(s_{\kappa_1+1}, \ldots, s_{\kappa_2}) \otimes \cdots \otimes M_{i-\kappa_{j-1}}(s_{\kappa_{j-1}+1}, \ldots, s_i)\big] \Bigg\}. \end{aligned} \tag{8}$$

A nice feature of these equations is their recursive nature. Once $M_1(s)$ has been found, it can be used to obtain an expression for $M_2(s_1, s_2)$, and so forth. Under partial linearization, the operators $E_k = 0$, $k > 1$. Assuming $E_1(s)$ and $P_1(s)$ to be invertible, then the kernels of the controller G are given by $G_1(s) = P_1^{-1}(s)T_1(s)E_1^{-1}(s)$,

$$
\begin{aligned}
&G_2(s_1,s_2) = -P_1^{-1}(s_1+s_2)P_2(s_1,s_2)\Big[G_1(s_1)\otimes G_1(s_2)\Big],\\
&G_3(s_1,s_2,s_3) = -P_1^{-1}(s_1+s_2+s_3)\Big\{P_2(s_1,s_2+s_3)\Big[G_1(s_1)\otimes G_2(s_2,s_3)\Big]\\
&+P_2(s_1+s_2,s_3)\Big[G_2(s_1,s_2)\otimes G_1(s_3)\Big]+P_3(s_1,s_2,s_3)\Big[G_1(s_1)\otimes G_1(s_2)\otimes G_1(s_3)\Big]\Big\}, \qquad (9)
\end{aligned}
$$

$$
\begin{aligned}
G_i(s_1,\ldots,s_i) &= -P_1^{-1}(s_1+\cdots+s_i)\\
&\sum_{j=2}^{i}\left\{\sum_{k_1=1}^{i-j+1}\ \sum_{k_2=1}^{i-j-k_1+2}\cdots\sum_{k_{j-1}=1}^{i-\kappa_{j-2}-1} P_j(s_1+\cdots+s_{\kappa_1},s_{\kappa_1+1}+\cdots+s_{\kappa_2},\ldots,s_{\kappa_{j-1}+1}+\cdots+s_i)\right.\\
&\left.\Big[G_{k_1}(s_1,\ldots,s_{\kappa_1})\otimes G_{k_2}(s_{k_1+1},\ldots,s_{\kappa_2})\otimes\cdots\otimes G_{i-\kappa_{j-1}}(s_{\kappa_{j-1}-1},\ldots,s_i)\Big]\right\}. \qquad (10)
\end{aligned}
$$

Again, note the recursive nature of the controller design equations. Once $G_1(s)$ has been computed, it can be used to find $G_2(s_1,s_2)$, and so on.

The next step in the realization process is to substitute the expressions for the plant kernels P_i in the frequency domain. Given $\overline{G}_i^{\,j} = G_i(s_j,s_{j+1},\ldots,s_{j+i-1})$, then

$$
\overline{G}_2^{\,1} = -\left(C_1\overline{P}_1^{\,2}\right)^{-1} C_2\overline{P}_2^{\,2}\left\{\left[\overline{P}_1^{\,1}\overline{G}_1^{\,1}\right]\otimes\overline{G}_1^{\,2}\right\}, \qquad (11)
$$

$$
\begin{aligned}
\overline{G}_3^{\,1} = &-\left(C_1\overline{P}_1^{\,3}\right)^{-1}\left\{C_2\overline{P}_2^{\,3}\left\{\left[\overline{P}_1^{\,1}\overline{G}_1^{\,1}\right]\otimes\overline{G}_2^{\,2}\right\}+C_2\overline{P}_2^{\,3}\left\{\left[\overline{P}_1^{\,2}\overline{G}_2^{\,1}\right]\otimes\overline{G}_1^{\,3}\right\}\right.\\
&\left.+C_3\overline{P}_3^{\,3}\left\{\left[\overline{P}_2^{\,2}\left\{\left[\overline{P}_1^{\,1}\overline{G}_1^{\,1}\right]\otimes\overline{G}_1^{\,2}\right\}\right]\otimes\overline{G}_1^{\,3}\right\}\right\}. \qquad (12)
\end{aligned}
$$

$$
\begin{aligned}
\overline{G}_i^{\,1} = -\left(C_1\overline{P}_1^{\,i}\right)^{-1}\sum_{j=2}^{i}&\left\{\sum_{k_1=1}^{i-j+1}\ \sum_{k_2=1}^{i-j-\kappa_1+2}\ \sum_{k_3=1}^{i-j-\kappa_2+3}\cdots\sum_{k_{j-1}=1}^{i-\kappa_{j-2}-1}\right. \qquad (13)\\
&\left.C_j\overline{P}_j^{\,i}\Big\{\Big[\overline{P}_{j-1}^{\,\kappa_{j-1}}\Big\{\Big[\cdots\Big[\overline{P}_2^{\,\kappa_2}\Big\{\Big[\overline{P}_1^{\,\kappa_1}\overline{G}_{k_1}^{\,1}\Big]\otimes\overline{G}_{k_2}^{\,\kappa_1+1}\Big\}\Big]\cdots\Big]\otimes\overline{G}_{k_{j-1}}^{\,\kappa_{j-2}+1}\Big\}\Big]\otimes\overline{G}_{i-\kappa_{j-1}}^{\,\kappa_{j-1}+1}\Big\}\right\}.
\end{aligned}
$$

Each of Eqs. (11)–(13) can be realized as a set of interconnected linear subsystems. These realizations possess a number of interesting features that are nicely illustrated using block diagrams. In doing so, certain notational conveniences are employed.

First, The superscripts on the quantities $\overline{P}_i^{\,j}$ and $\overline{G}_i^{\,j}$ in Eqs. (11)–(13) are redundant, and can be neglected without loss [9]. The proof lies in observing that the ordering of terms in a controller design equation, together with the established use of parentheses and brackets, are sufficient to determine what the superscripts should be.

Second, if $e \in \mathbb{R}^p$ represents the output error of the closed loop, and δ_{11} is its Laplace transform, then let $\delta_{11}^{(i)}$ denote the i-fold tensor product of δ_{11} with itself, and define $\beta_{i1} = \sum_{k=1}^{i}\overline{G}_k^{\,1}(\delta_{11}^{(k)})$, making β_{i1} the Laplace transform of the ith order controller output.

Finally, the layout of the block diagrams follows certain conventions. If the controller design equations are fully expanded and each written on one line, then their addends, as read from left to right, appear in the block diagrams proceeding from top to bottom.

The fundamental observation behind the simplification process is that quantities present in the realization for $\overline{G}_i^{\,1}$ reappear in the realizations for $\overline{G}_j^{\,1}$, $j > i$. The motivation behind the simplification process is that in computing the output of a given controller component $\overline{G}_i^{\,1}$, the most efficient algorithm is to compute these repeated quantities only once, store them in memory, and then reuse them as needed. The reduction process is thus a problem of identifying a set Γ of elements common to a sequence of controller components $\overline{G}_i^{\,1}$, $i = 1,2,3,\ldots,n$, and determining which ones to save for later reuse.

In computing the n components $\overline{G}_i^1(\delta_{11}^{(i)})$, $1 \le i \le n$, of the output of an nth order controller, many quantities are computed repeatedly. Therefore, a system of notation is now introduced with the primary goal of eliminating redundant calculations by identifying members of the common set Γ, and the secondary goal of further simplifying the controller output calculation. The identification of Γ presented here is not unique. Therefore, the notation Γ_0 will be used to denote the particular identification described herein, and to distinguish it from the general notion of the common set represented by Γ. Here and in the sequel, superscripts are suppressed.

The first members of Γ_0 have already been introduced, namely, β_{i1}. Given $\alpha_{ij} \in \Gamma_0$, $2 \le j \le i$, defined by $\alpha_{ij} = \overline{P}_{j-1}\beta_{(i-1)(j-1)}$, then [9] $\beta_{ij} = \sum_{k=1}^{i-j+1} \alpha_{(i-k+1)j} \otimes \beta_{k1}$.

The simplified partial linearization nth order controller output can now be calculated as follows. First, compute the output of the first order controller, $\beta_{11} = P_1^{-1}T_1(I - T_1)^{-1}\delta_{11}$. Then, the output of the nth order controller, $n \ge 2$, is computed using

$$\beta_{n1} = \beta_{11} - \sum_{i=2}^{n}\sum_{j=2}^{i} P_1^{-1}C_j\overline{P}_j\beta_{ij}. \tag{14}$$

The members of the common set Γ_0, (namely, α_{ij} and β_{ij}), are computed and stored in memory the first time they are encountered, and then recalled as needed. In the sequel, the following identifications are convenient: $g_{ij} = -P_1^{-1}C_j\overline{P}_j\beta_{ij}$, and

$$g_i = \sum_{j=2}^{i} -P_1^{-1}C_j\overline{P}_j\beta_{ij} = \sum_{j=2}^{i} g_{ij}. \tag{15}$$

The above algorithm for calculating the output of the simplified partial linearization controller takes full advantage of the recursive nature of the design equations, and provides a significant improvement over the procedure obtained via a straightforward application of the general controller realizations given by Al-Baiyat [6], in that the number of states used in the realization is greatly reduced, as are floating-point computational operations. Furthermore, the simplified form is amenable to the construction of general software routines, capable of computing the controller output for an arbitrary value of the controller order n.

The simplified block diagrams for the second third order controller components are shown in Fig.s (3)–(4). A general form for the ith order component is shown in Fig. (5).

6 Single Degree of Freedom Base-Isolated Structure Example

Consider an SDOF structure sitting on a base isolation system consisting of hysteretically-damped bearings, and let the structure be subject to ground acceleration $\ddot{x}_g(t)$ and a control force $f(t)$ supplied by a hydraulic actuator. Such a system is shown in Fig. (6). The forces exerted by the individual bearings are modeled collectively. The equation of motion for the horizontal displacement of the mass m of the structure is $m\left[\ddot{x}_p(t) + \ddot{x}_g(t)\right] + c\dot{x}_p(t) + kx_p(t) + Q[x(t), \dot{x}(t)] = f(t)$. Descriptions of the variables and parameter values in Eq. (sdofeom) are given in Table (1). Both the bearings and the actuator affect the stiffness and damping coeffecients c and k in Eq. (sdofeom), and it should be understood that the values identified for these parameters provided in Table (1) are meant to represent the behavior of the system with the actuator and bearings in place. The hysteretic restoring force Q generated by the base isolating bearings is given by $Q[x_p(t), \dot{x}_p(t)] = \alpha\frac{F_y}{Y}x_p(t) + (1-\alpha)F_y z(t)$, where $z(t)$ is a dimensionless quantity representing hysteretic displacement [12, 13, p. 2-6] satisfying the first order differential

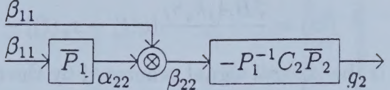

β_{11}
β_{11}
$\overline{P}_1$
α_{22}
β_{22}
$-P_1^{-1}C_2\overline{P}_2$
g_2

Figure 3: Simplified realization of G_2

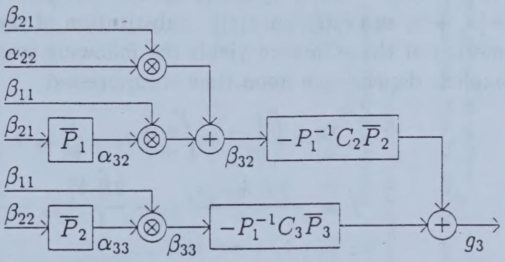

β_{21}
α_{22}
β_{11}
β_{21}
$\overline{P}_1$
α_{32}
β_{32}
$-P_1^{-1}C_2\overline{P}_2$
β_{11}
β_{22}
$\overline{P}_2$
α_{33}
β_{33}
$-P_1^{-1}C_3\overline{P}_3$
g_3

Figure 4: Simplified realization of G_3

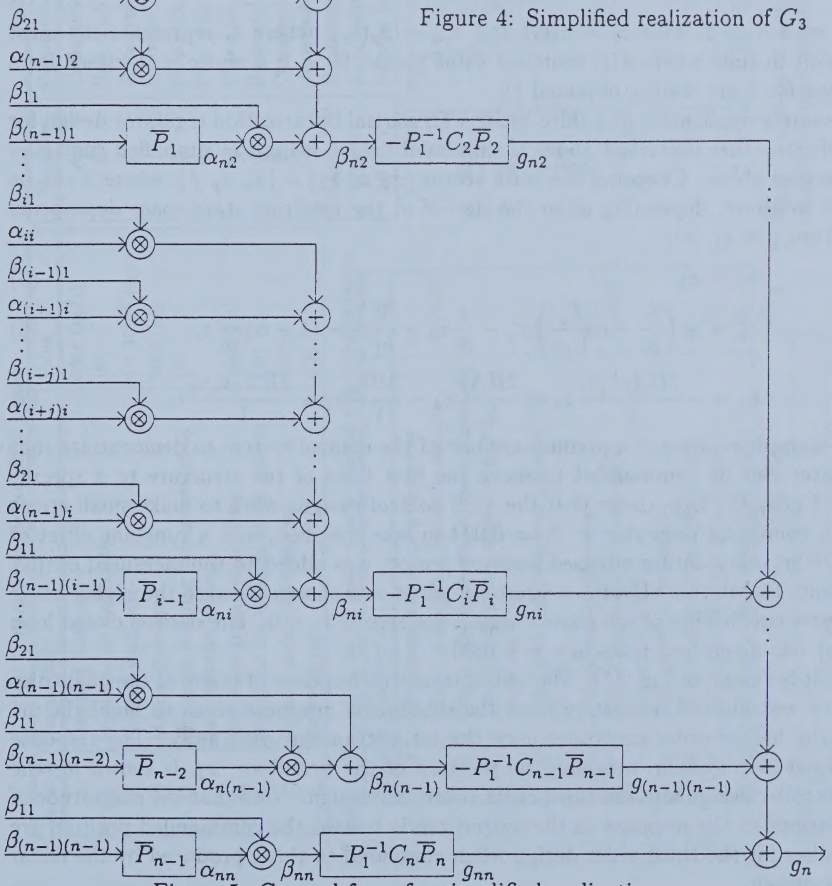

$\beta_{(n-1)1}$
α_{22}
$\beta_{(n-2)1}$
α_{32}
β_{31}
$\alpha_{(n-2)2}$
β_{21}
$\alpha_{(n-1)2}$
β_{11}
$\beta_{(n-1)1}$
$\overline{P}_1$
α_{n2}
β_{n2}
$-P_1^{-1}C_2\overline{P}_2$
g_{n2}
β_{i1}
α_{ii}
$\beta_{(i-1)1}$
$\alpha_{(i+1)i}$
$\beta_{(i-j)1}$
$\alpha_{(i+j)i}$
β_{21}
$\alpha_{(n-1)i}$
β_{11}
$\beta_{(n-1)(i-1)}$
$\overline{P}_{i-1}$
α_{ni}
β_{ni}
$-P_1^{-1}C_i\overline{P}_i$
g_{ni}
β_{21}
$\alpha_{(n-1)(n-1)}$
β_{11}
$\beta_{(n-1)(n-2)}$
$\overline{P}_{n-2}$
$\alpha_{n(n-1)}$
$\beta_{n(n-1)}$
$-P_1^{-1}C_{n-1}\overline{P}_{n-1}$
$g_{(n-1)(n-1)}$
β_{11}
$\beta_{(n-1)(n-1)}$
$\overline{P}_{n-1}$
α_{nn}
β_{nn}
$-P_1^{-1}C_n\overline{P}_n$
g_{nn}
g_n

Figure 5: General form for simplified realization.

equation $Y\dot{z}(t) = \gamma|\dot{x}_p(t)|\,|z(t)|^{n-1}z(t) - \beta\dot{x}_p(t)|z(t)|^n + A\dot{x}_p(t)$. Table (2) lists descriptions of the variables and parameter values. The actuator force f applied to the SDOF structure is governed by the equation [14, 15]

$$\dot{f}(t) = \frac{2BA_f k_q \gamma_f}{V}[u(t) - x_p(t)] - \frac{2Bk_c}{V}f(t) - \frac{2BA_f^2}{V}\dot{x}_p(t), \tag{1}$$

where $u(t)$ is the control signal generated by the controller. Descriptions of the quantities in Eq. (1) are provided in Table (3) [16].

Defining $A_z = \frac{A}{Y}$, $\beta_z = \frac{\beta}{Y}$ and $\gamma_z = \frac{\gamma}{Y}$, and using the fact that $|z| = z \operatorname{sgn} z$, then $\dot{z}(t) = [A_z - (\beta_z + \gamma_z \operatorname{sgn}\dot{x}_p \operatorname{sgn} z)z^n(t)]\dot{x}_p(t) = [A_z + \sigma(t)z^n(t)]\dot{x}_p(t)$, where $\sigma(t) = -[\beta_z + \gamma_z \operatorname{sgn}\dot{x}_p(t) \operatorname{sgn} z(t)]$. Substitution of the expression for Q into the equation of motion of the structure yields the following system of equations. For convenience, the explicit dependence upon time is suppressed.

$$\ddot{x}_p = -\left(\frac{k}{m} + \alpha\frac{F_y}{Ym}\right)x_p - \frac{c}{m}\dot{x}_p + \frac{1}{m}f(t) - (1-\alpha)\frac{F_y}{m}z - \ddot{x}_g, \tag{2}$$

$$\dot{f} = -\frac{2BA_f k_q \gamma_f}{V}x_p - \frac{2BA_f^2}{V}\dot{x}_p - \frac{2Bk_c}{V}f(t) + \frac{2BA_f k_q \gamma_f}{V}u, \tag{3}$$

$$\dot{z} = (A_z + \sigma z^n)\dot{x}_p. \tag{4}$$

Given z with $n = 2$, with $z_0 = z(t_\sigma)$ and $x_{p0} = x_p(t_\sigma)$, where t_σ represents the most recent point in time where $\sigma(t)$ switched value yields, for $\sigma \geq 0$ and $\sigma < 0$, closed-form expressions for z are readily obtained [9].

An example application of a third order VFS partial linearization regulator design for the SDOF structure described above is simulated below, using the simplified controller design derived above. Choosing the state vector $[\,x_1\ x_2\ x_3\,] = [\,x_p\ \dot{x}_p\ f\,]$, where z can be expressed as above, depending upon the sign of σ, the resulting state space description has the form $y = x_1$, and

$$\dot{x}_1 = x_2 \tag{5}$$

$$\dot{x}_2 = -\left(\frac{k}{m} + \alpha\frac{F_y}{Ym}\right)x_1 - \frac{c}{m}x_2 + \frac{1}{m}x_3 - (1-\alpha)\frac{F_y}{m}z, \tag{6}$$

$$\dot{x}_3 = -\frac{2BA_f k_q \gamma_f}{V}x_1 - \frac{2BA_f^2}{V}x_2 - \frac{2Bk_c}{V}x_3 + \frac{2BA_f k_q \gamma_f}{V}u, \tag{7}$$

This example represents a preliminary test of the control system to demonstrate that the actuator can be commanded to move the first floor of the structure to a specific location. Under the hypothesis that the VFS control designs work to make small errors smaller, a command request $r = \overline{x}_1 = 0.004$ m was specified, and a constant offset of $y_0 = 0.001$ m, representing a biased position sensor, was added to the measured output of the plant. Regulation about a nonzero set point was chosen because the higher order Taylor series coefficients of the control model are zero if $\overline{x}_1 = 0$. The desired closed loop map $T_1(s)$ was given four poles at $s = -1000$.

A result is shown in Fig. (7). The initial transient response of many of the quantities that follow was plotted separately from the subsequent response so as to highlight the effect of the higher order controller once the intial transient died away. The response of the closed-loop system, meaning the position of the first floor, x_1, is shown for the linear controller design and the third order controller design. Note that the magnitude of the oscillations in the response as the output tends toward the commanded position are much smaller for the third order design when compared to those produced by the linear controller design.

Table 1: Parameters for the SDOF system

Variable	Description	Value
c	Damping coefficient	600π (kg/s)
f	Force applied by actuator	(N)
k	Stiffness coefficient	$3 \times 10^6\pi^2$ (kg/s^2)
m	Mass of structure	3000 (kg)
Q	Hysteretic restoring force generated by base isolation bearing	(N)
x_g	Horizontal ground acceleration	(m/s^2)
x_p	Horizontal displacement of the mass m	(m)
f_p	Natural undamped frequency	5 (Hz)
ξ_p	Damping ratio	1 %

Table 2: Parameters for the hysteretically damped base isolation system

Variable	Description	Value
A	Hysteresis loop-shaping parameter	1
F_y	Force of equivalent hysteretic damper	40 (KN)
n	Parameter controlling the smoothness of the transition from elastic to plastic response	2
Q	Hysteretic restoring force generated by bearing	(N)
x_p	Horizontal displacement	(m)
Y	Yield displacement of equivalent hysteretic damper	1 (mm)
z	Hysteretic displacement	
α	Post to preyielding stiffness ratio	0.2
β	Hysteresis loop-shaping parameter	-0.25
γ	Hysteresis loop-shaping parameter	0.75

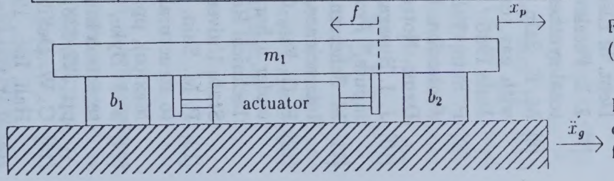


Table 3: Parameters for the actuator

Variable	Description	Value
A_f	Cross-sectional area of actuator	(m^2)
B	Bulk modulus of hydraulic fluid	(N/m^2)
f	Force applied to SDOF structure by actuator	(N)
g	Proportional feedback gain	2.5
k_c	Flow-pressure coefficient	(m^4·s/kg)
k_q	Flow gain	(m^2/s)
x_p	Horizontal displacement	(m)
V	Characteristic hydraulic fluid volume for the actuator	(m^3)
$2BA_fk_qG/V$	Displacement coefficient	3.6484×10^8 (N/m/s)
$2Bk_c/V$	Force feedback coefficient	66.67 (1/s)
$2BA^2/V$	Velocity feedback coefficient	2.1891×10^7 (N/m)

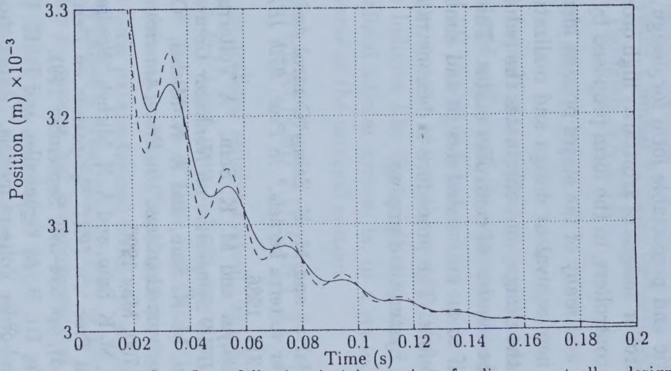


Figure 7: Position of first floor following intial transient for linear controller design (dashed) and third order controller design (solid).

Figure 6: An SDOF structure of mass m_1 with base isolation provided by the hysteretically-damped bearings b_i, $i = 1, 2$, subject to ground acceleration $\ddot{x}_g(t)$ and a control force $f(t)$.

7 Conclusion

A systematic and predictable form for design and realization of Volterra partial linearization controller designs of arbitrarily high order is obtained for class of linear analytic plants. These controllers, in the form proposed by Al-Baiyat, possess a significant degree of complexity, and many of the states possess identical dynamic behavior, indicating redundancy. A new equivalent design and realization is derived that greatly reduces the complexity of the design, and eliminates the redundant dynamic behavior, permitting a reduction in the number of controller states. The simplified form is recursive in nature, and lends itself to an automated design and simulation procedure, something that was not possible before. The procedure is demonstrated on a regulator design for an SDOF structure with hysteretic damping.

References

[1] S. A. Al-Baiyat and M. K. Sain, "Control design with transfer functions associated to higher-order volterra kernels," in *Proc. 25th IEEE Conference on Decision and Control*, pp. 1306–1311, 1986.

[2] S. A. Al-Baiyat and M. K. Sain, "A Volterra method for nonlinear control design," *Preprints IFAC Symposium on Nonlinear Control System Design*, pp. 76–81, June 1989.

[3] P. M. Sain, M. K. Sain, and A. N. Michel, "On coordinated feedforward excitation of nonlinear servomechanisms," in *Proc. American Control Conf.*, (San Diego, California), pp. 1695–1700, May 1990.

[4] P. M. Sain, M. K. Sain, and A. N. Michel, "Nonlinear model-matching design of servomechanisms," in *Proc. 1st IFAC Symposium on Design Methods of Control Systems*, (Zurich, Switzerland), pp. 594–599, September 1991.

[5] F. J. Doyle, III, B. A. Ogunnaike, and R. K. Pearson, "Nonlinear model-based control using second-order Volterra models," *Automatica*, vol. 31, pp. 697–714, May 1995.

[6] S. A. Al-Baiyat, *Nonlinear Feedback Synthesis: A Volterra Approach.* Ph.D. Dissertation, Dept. of Electrical Engineering, Univ. of Notre Dame, Notre Dame, Indiana, May 1986.

[7] J. L. Peczkowski and M. K. Sain, "Design of nonlinear multivariable feedback controls by total synthesis," in *Proc. 1984 American Control Conf.*, pp. 688–697, 1984.

[8] M. K. Sain and J. L. Peczkowski, "Nonlinear control by coordinated feedback synthesis with gas turbine applications," in *Proc. 1985 American Control Conf.*, pp. 1121–1128, June 1985.

[9] P. Sain, *Volterra Control Synthesis, Hysteresis Models and Magnetorheological Structure Protection.* Ph.D. Dissertation, Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana, 46556, May 1997.

[10] J. J. Bussgang, L. Ehrman, and J. Graham, "Analysis of nonlinear systems with multiple inputs," *Proc. IEEE*, vol. 62, pp. 1088–1119, Aug. 1974.

[11] M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems.* New York: Wiley Interscience, 1980.

[12] F.-G. Fan, G. Ahmadi, and I. G. Tadjbakhsh, "Base isolation of a multi-story building under a harmonic ground motion — a comparison of performances of various systems," Technical Report NCEER-88-0010, National Center for Earthquake Engineering Research, State University of New York at Buffalo, May 1988.

[13] P. M. Sain, M. K. Sain, and B. F. Spencer, Jr., "Models for hysteresis and application to structural control," in *Proc. 1997 American Control Conf.*, vol. 1, (Albuquerque, New Mexico), pp. 16–20, June 1997.

[14] S. J. Dyke, B. F. Spencer, Jr., P. Quast, and M. K. Sain, "Role of control-structure interaction in protective system design," *ASCE Journal of Engineering Mechanics*, vol. 121, pp. 322–338, Feb. 1995.

[15] C. W. DeSilva, *Control Sensors and Actuators.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1989.

[16] S. J. Dyke, "Actuator model notes." Personal communication. Univ. of Notre Dame, Notre Dame, Indiana, Oct. 1995.

GENERALIZED ADJOINT SYSTEMS INVERSES

Demetrios Serakos
Naval Surface Warfare Center, Combat Systems Department
Code N24, Dahlgren, VA 22448

1. INTRODUCTION

A representation for the inverse of a causal input-output system is given in this paper. This inverse is in terms of the generalized adjoint of the system. Generalized adjoint systems have been discussed in [1] and [2]. Another input-output system, referred to as the auxiliary system, which has as input/output spaces the output/input spaces of the generalized adjoint system and which behaves like the original system, is defined and described. It is shown that the generalized adjoint system is, in a sense, the inverse of the auxiliary system and hence the inverse of the original system.

2. REVIEW OF THE GENERALIZED ADJOINT SYSTEM

This section gives an outline of a generalized adjoint system, which is discussed in [1] and [2]. An input-output system (Y, F, U) is a map F from an input space U to an output space Y, $F : U \to Y$. U and Y are spaces of time functions, U is closed with respect to translations. A family of seminorms is defined on U and Y. For U, this is $\mathcal{N}_U = \{\|\cdot\|_{s,t}, -\infty < s < t < \infty\}$. If this time parameterized family of seminorms satisfies conditions related to time intervals, it is called a fitted family (FF) of seminorms, see [1] and [3]. FF's work similarly to standard L_p spaces, but are more general, e.g., weighted L_p spaces are included. In a sense $\|u\|_t \triangleq \|u\|_{-\infty,t}$ and $\|u\| \triangleq \|u\|_{-\infty,\infty}$. Here the system (Y, F, U) is causal if for all $u^1, u^2 \in U$ and all t, $\|u^1 - u^2\|_t = 0$ gives $\|F(u^1) - F(u^2)\|_t = 0$. (Y, F, U) has memory length $M > 0$ if $\|u^1 - u^2\|_{s-M,t} = 0$ then $\|F(u^1) - F(u^2)\|_{s,t} = 0$. If $M = 0$ works, the system is memoryless. If (Y, F, U) is causal, a truncated system $F_t : U_t \to Y_t$ may be defined and $t \to F_t$ is a time function. Define a seminorm on F_t by $[\![F_t]\!]_{s,t} \triangleq \sup_{u \in U} \|F(u)\|_{s,t} / (1 + \|u\|_t^N)$. Conditions are given in [1] so that $\{[\![\cdot]\!]_{s,t}\}$ form a FF of seminorms for $\{(Y, F^i, U)\}$. Consider the spaces $\mathcal{B} = \{(Z, B^i, Y)\}$ and $\mathcal{A} = \{(Z, A^i, U)\}$. View these as an input space, $Y^a = \mathcal{B}$, and output space $U^a = \mathcal{A}$. Translation for $B \in Y^a$ is defined by $\mathrm{L}^*_\tau B = \mathrm{L}_\tau B \mathrm{R}_\tau$, see [1]. The generalized adjoint of a causal input-output system (Y, F, U) is the system (U^a, F^a, Y^a); $F^a : Y^a \to U^a$ defined by $F^a(B) \triangleq BF = A$. For a set of causal input-output systems $\mathcal{F} = \{(Y, F^i, U)\}$, $\mathcal{F}^a = \{(U^a, F^{ia}, Y^a)\}$ is the set of adjoint systems. An adjoint map $\phi : \mathcal{F} \to \mathcal{F}^a$ is defined by $\phi(F) = F^a$. Properties of F^a include F^a is linear and memoryless (hence causal). Conditions are given for F^a to be bounded and ϕ to be linear, bounded, preserve translations and have a bounded inverse. If $B \in Y^a$ are Lipschitz continuous, ϕ is continuous.

3. SUMMARY OF RESULTS

3.1. Invertability For F and F^a

In the first place, we assume that F and F^a are invertible and find a relationship between the inverses. From the definitions, $F^a(B) = A = BF$. However, since F is invertible, $AF^{-1} = B$; hence, $(F^{-1})^a A = B$. Substituting, we obtain $(F^{-1})^a F^a(B) = B$. Substituting again, we obtain $F^a(F^{-1})^a A = A$ and conclude that $(F^a)^{-1} = (F^{-1})^a$.

We next consider the invertibility of F^a based on the invertibility of F, in the algebraic sense. That is, we consider whether F^a is 1 to 1 and onto if F is 1 to 1 and onto. Similar analysis for the invertibility of F based on F^a is also considered.

(3.1.A). Assume that F is onto. Let B_1, $B_2 \in Y^a$ such that $B_1 \neq B_2$, that is, $B_1(y) \neq B_2(y)$ for some $y \in Y$. Pick $u \in U$ such that $F(u) = y$. This is possible since F is assumed to be onto. Then, $A_1(u) = [F^a(B_1)](u) = B_1F(u) = B_1(y)$ and $A_2(u) = [F^a(B_2)](u) = B_2F(u) = B_2(y)$. Hence $A_1 \neq A_2$, and so F onto implies F^a is 1 to 1.

(3.1.B). It may be shown that F is 1 to 1 implies that F^a is onto.

The complement of these may also be obtained. These are: F^a onto implies F 1 to 1 (also requires for every $u_1 \neq u_2$ there exists A such that $A(u_1) \neq A(u_2)$, see LEMMA 4); and F^a 1 to 1 implies F onto (also requires for every $\hat{y} \in Y$ there exist $B_1, B_2 \in Y^a$ such that $B_1(y) = B_2(y)$ for all $y \in Y - \{\hat{y}\}$, and that $B_1(\hat{y}) \neq B_2(\hat{y})$).

3.2. ON F^a AS THE INVERSE OF F

The capacity in which F^a is the inverse of F is considered. Since the domain and range of F is not the same as, respectively, the range and domain of F^a, the first part of this section is to construct the auxiliary system. In this section, let $\mathcal{N}_Y = \{\|\cdot\|_{s,t}, s < t\}$ be a FF of seminorms defined on Y with $\alpha_Y = +\infty$ and $I\!K_Y = 1$. Likewise, $\mathcal{N}_U$ is defined on U with $\alpha_U = +\infty$ and $I\!K_U = 1$. The "sup"-norm (L_∞ norm) is used on Z.

(3.2.A). An A is constructed which has specialized properties. Consider a $z \in Z$; z is (in general) an equivalence class of time functions. Let $[z]$ denote this equivalence class and let $\zeta \in [z]$ be a specified element. For each $u^\# \in U$, specify an $A^\# \in U^a$ by

$$A_t^\#(u_t)(t) = \left\|u_t - u_t^\#\right\|_t \cdot \zeta(t) \ . \tag{1}$$

The set of all such $A^\#$ is $\{A^\#\} = \mathcal{A}^\# \subset \mathcal{A}$. The map $\gamma : U \to \mathcal{A}^\#$ defined by $\gamma(u) = A^\#$ is one to one and onto. In fact, γ is onto by construction. Several properties of the systems $A^\#$ are considered. It may be shown that $A^\#$ is causal, bounded and continuous. Properties of γ include boundedness and uniform continuity. If we let $\inf_t \|\zeta(t)\| > 0$, the map $A^\# \to u^\#$, (γ^{-1}), is uniformly continuous and bounded.

CLOSURE OF $\mathcal{A}^\#$ WITH RESPECT TO TRANSLATION: (Note that $\mathrm{L}_\tau \|u\|_t = \|\mathrm{L}_\tau u\|_t$ and that $\mathrm{L}_\tau \|u_t\|_t = \|u_{t+\tau}\|_t$.) Hence,

$$\mathrm{L}_\tau^* A^\#(u)(t) = \mathrm{L}_\tau(A^\#(\mathrm{R}_\tau u)(t)) = \left\|u_t - u_{t+\tau}^\#\right\|_t \cdot \zeta(t+\tau) \ . \tag{2}$$

We consider the effect of a translation in U: Let $\mathrm{L}_\tau u^\# = u^\flat$, then

$$A^\flat(u)(t) = \left\|u_t - (\mathrm{L}_\tau u^\#)_t\right\|_t \cdot \zeta(t) = \left\|u_t - {u^\#}_{t+\tau}\right\|_t \cdot \zeta(t) \ . \tag{3}$$

Let "$\overset{map}{\rightarrow}$" represent the map γ. Using (2) we get $u^\# \overset{map}{\rightarrow} A^\# \overset{\mathrm{L}_\tau^*}{\rightarrow} (\mathrm{L}_\tau^* A^\#)$. Using (3) we get $u^\# \overset{\mathrm{L}_\tau}{\rightarrow} \mathrm{L}_\tau u^\# = u^\flat \overset{map}{\rightarrow} A^\flat$. These do not form a commutative diagram since (2) does not equal (3), in general. If, however, ζ is selected such that $\zeta(t+\tau) = \zeta(t)$, then a commutative diagram is formed. With the norm selected on Z, $\zeta(t+\tau) = \zeta(t)$ is possible. To be viewed as an input space, $\mathcal{A}^\#$ will have to be closed with respect to translation; hence, may have to be expanded.

(3.2.B). As stated above, a map which has the same input and output spaces as the output and input spaces of F^a, and which behaves similarly to F is needed. In this section a map G is defined to fulfill this need. Define a map $G : \mathcal{A}^\# \to \mathcal{B}^\#$ by

$$G(A^\#)y_t(t) \triangleq \left\|y_t - F_t(u_t^\#)\right\|_t \cdot \zeta(t) \ . \tag{4}$$

The map G is related to F in the following sense. If $y^\# = F(u^\#)$ and $A^\# \in \mathcal{A}^\#$, $B^\# \in \mathcal{B}^\#$ are derived from $u^\#$ and $y^\#$ respectively as described above, then $G(A^\#) = B^\#$.

INVERTIBILITY: Next, we consider whether F^a inverts G. Assume that F is 1 to 1.

$$(F^a(B^\#))_t u_t(t) = B_t^\#(F_t u_t)(t) = \left\|F_t u_t - y_t^\#\right\|_t \cdot \zeta(t)$$
$$= \left\|F_t u_t - F u_t^\#\right\|_t \cdot \zeta(t) \equiv \left\|u_t - u_t^\#\right\|_t \cdot \zeta(t) = (A_t^\#) u_t(t) \ . \qquad (5)$$

The equivalence relation "$\equiv$" is based on null spaces. Hence, we get from (4) and (5), $F^a(G(A^\#)) \equiv A^\#$.[1]

MEMORY CONSIDERATIONS: It is shown that G is memoryless. Consider $A^\#$, $A^\flat$ such that $[\![A^\# - A^\flat]\!]_{s,t} = 0$ for all $s, t \in \Re$. This gives that for all $u \in U$, $\left\|A^\#(u) - A^\flat(u)\right\|_{s,t} = \left\|u_\tau - u_\tau^\#\right\|_\tau \cdot \zeta(\tau) - \left\|u_\tau - u_\tau^\flat\right\|_\tau \cdot \zeta(\tau) = 0$, for all $s < \tau \le t$. Substituting $u = u^\#$in gives $\left\|u_\tau^\# - u_\tau^\flat\right\|_\tau \cdot \zeta(\tau) = 0$. Now, for any $y \in Y$the causality of F gives

$$[G(A^\#) - G(A^\flat)](y)(\tau) \le \left\|F_\tau(u_\tau^\flat) - F_\tau(u_\tau^\#)\right\|_\tau \cdot \zeta(\tau) = 0$$

for all $s < \tau \le t$. This implies $[\![G(A^\#) - G(A^\flat)]\!]_{s,t} = 0$ which indicates that G is memoryless. One viewpoint of how a linear map, F^a, could be the inverse of a nonlinear map, G, is that the input space of G is not a linear space, in fact, it is not closed with respect to addition.

BOUNDEDNESS OF G:

$$[\![G]\!] \triangleq \sup_t [\![G_t]\!]_t = \sup_t \sup_{A^\# \in \mathcal{A}^\#} \frac{[\![G_t(A_t^\#)]\!]_t}{1 + [\![A_t^\#]\!]_t^N} \le \sup_{t,y,A^\#} \frac{\left\|y_t - F_t(u_t^\#)\right\|_t}{(1 + \|y_t\|_t^N)(1 + [\![A_t^\#]\!]_t^N)} \cdot \sup_t \|\zeta(t)\| \ .$$

Assume that $\|\zeta(t)\| = 1$ for all t. Then, $[\![G]\!] \le 1 + \sup_{t, A^\# \in \mathcal{A}^\#} \left\|F_t(u_t^\#)\right\|_t / (1 + [\![A_t^\#]\!]_t^N)$. A lower bound for $[\![A_t^\#]\!]_t$ is needed. Consider

$$[\![A_t^\#]\!]_t = \sup_u \frac{\sup_{\tau \le t} \left\|u_\tau - u_\tau^\#\right\|_\tau \cdot \|\zeta(\tau)\|}{1 + \|u_t\|_t^N} \ge \sup_u \frac{\left|\|u_t\|_t - \left\|u_t^\#\right\|_t\right|}{1 + \|u_t\|_t^N} \ge \left\|u_t^\#\right\|_t \ .$$

Substituting in the above gives $[\![G]\!] \le 1 + \sup_t [\![F_t]\!] = 1 + [\![F]\!]$.

CONTINUITY OF G: F_t Lipschitz continuity is needed.

$$[\![G(A^\#) - G(A^\flat)]\!] = \sup_{t,y} \left(\left(\left\|G_t(A_t^\#)(y_t) - G_t(A_t^\flat)(y_t)\right\|_t \right) / (1 + \|y_t\|_t^N) \right)$$

$$\le \sup_t \left\|F_t(u_t^\#) - F_t(u_t^\flat)\right\|_t \cdot \sup_t \|\zeta(t)\| \le K \cdot \sup_t \left\|u_t^\# - u_t^\flat\right\|_t \cdot \sup_t \|\zeta(t)\|$$

where K is the Lipschitz constant. Hence, G is continuous (using the continuity of γ^{-1}).

(3.2.C). In this section the invertibility of the adjoint map is considered. The adjoint map is the map $\phi : \mathcal{F} \to \mathcal{F}^a = \{(U^a, F^{ia}, Y^a)\}$. The Lipschitz norm of ϕ is $[\![\phi]\!]_{Lip} = \sup_{F,H \in \mathcal{F}} [\![\phi(F) - \phi(H)]\!] / [\![F - H]\!]$. The next two Lemmas give conditions under which the Lipschitz norm is bounded below.

Lemma 1 *Under the condition that there exists a bounded linear* $B^I \in \mathcal{B}$, *for all* F, H $[\![\phi(F) - \phi(H)]\!] > (m/[\![B^I]\!]) \cdot [\![F - H]\!]$.

---

[1] Alternatives to (1) and (4) are: $A_t^\#(u_t)(t) = \delta(u_t - u_t^\#) \cdot \zeta(t)$ and $G(A^\#)(y_t)(t) = \delta(y_t - F_t(u_t^\#)) \cdot \zeta(t)$; where $\delta : U_t \to \Re$ is the delta function, that is $\delta(u_t) = 1$ if u_t is zero, and $= 0$ else. This $A^\#$ is bounded, but not continuous; however, the equivalence relation becomes an equality.

Lemma 2 *Under the condition that there exists bounded linear $B^* \in \mathcal{B}$ of the kind discussed above and in Appendix 4.1, for all F, H, $\llbracket \phi(F) - \phi(H) \rrbracket > \llbracket F - H \rrbracket$.*

Proof: Consider $\llbracket \phi(F) - \phi(H) \rrbracket = \sup_{t,B_0}(\llbracket (F^a_t - H^a_t)B_0 \rrbracket_0 / \llbracket B_0 \rrbracket_0) \geq \sup_t \llbracket B^*_t(F_t - H_t) \rrbracket_0 / \llbracket B^*_t \rrbracket_0$ for any particular (linear) B^*. Let B^* be of the form $B^*(y)(t) \triangleq f^t(y_t) \cdot \zeta^*(t)$ with $\llbracket f^t \rrbracket = 1$ and $\|\zeta^*(t)\| = 1$ for all t. Then, $B^*(y)(t)$

$$= \sup_t \left(\sup_{u_0} \frac{\|f^t((F_t - H_t)(u_0)) \cdot \zeta^*(t)\|}{1 + \|u_0\|_0^N} \Big/ \sup_{y_0} \frac{\|f^t(y_0) \cdot \zeta^*(t)\|}{\|y_0\|_0} \right) \geq \sup_{t,u_0} \frac{|f^t((F_t - H_t)(u_0))|}{1 + \|u_0\|_0^N} .$$

Now, for all $\epsilon > 0$, $\exists\, \bar{t}$ and v such that $\llbracket F - H \rrbracket \leq \llbracket F_{\bar{t}} - H_{\bar{t}} \rrbracket_0 + \epsilon/2 \leq (\|(F_{\bar{t}} - H_{\bar{t}})v_0\|_0)/(1 + \|v_0\|_0^N) + \epsilon$. Select, (using LEMMA 3), f^t such that $\llbracket f^t \rrbracket = 1$ and such that $f^t((F-H)_t v_0) = \|(F - H)_t v_0\|_0$. Then, $\llbracket \phi(F - H) \rrbracket + \epsilon \geq \|(F_{\bar{t}} - H_{\bar{t}})v_0\|_0 / (1 + \|v_0\|_0^N) + \epsilon \geq \llbracket F - H \rrbracket$. Since ϵ is arbitrary, $\llbracket \phi(F - H) \rrbracket \geq \llbracket F - H \rrbracket$. □

Both lemmas 1 and 2 give that ϕ is 1 to 1. Let $F, H \in \mathcal{F}$ with $F \neq H$. Then $\llbracket F - H \rrbracket > 0$. Both lemmas 1 and 2 imply $\llbracket \phi(F) - \phi(H) \rrbracket > 0$, which means $\phi(F) \neq \phi(H)$, so ϕ is one to one. Lemmas 1 and 2 also give that ϕ^{-1} is uniformly continuous.

APPENDIX

Let Y be the input space and Z be the output space. Fix specific $y^* \in Y$ and $z^* \in Z$. In general z^* is an equivalence class of time functions. Pick a time function ζ^* which is an element of z^*; that is, $\zeta^* \in [z^*]$. For each $t \in \Re$, there exists a linear functional $f^t : Y_t \to \Re$ such that $\llbracket f^t \rrbracket = 1$ for all t and such that $f^t(y^*_t) = \|y^*_t\|_t$ (using the Hahn-Banach theorem, [4]). Define an input-output system (Z, B, Y) by $(B^*(y))(t) \triangleq f^t(y_t)\zeta^*(t)$. Note that the output of B^* is a time function; however, the time function in turn specifies an equivalence class of time functions, which is an element of Z.

Lemma 3 *The input-output system B^* as defined is a causal linear system, and if the norm (L_∞) norm is used on Z it is a bounded input-output system.*

Let y^1 and y^2 be two distinct elements of Y. Let $\bar{t}$ be such that $y^1_{\bar{t}} \neq y^2_{\bar{t}}$. Define $y^* \triangleq y^1 - y^2$. Using LEMMA 3, and the input-output system defined there,

$$\|B^*(y^*)\| = \sup_t \left\| B^*(y^1) - B^*(y^2) \right\|_t \geq \left\| f^{\bar{t}}(y^1_{\bar{t}} - y^2_{\bar{t}}) \cdot \zeta^*(\bar{t}) \right\|_{\bar{t}} = \left\| y^1_{\bar{t}} - y^2_{\bar{t}} \right\| \cdot \zeta^*(\bar{t}) \neq 0 .$$

Hence, $B^*(y^1) \neq B^*(y^2)$. This result is restated in the following Lemma:

Lemma 4 *Given two distinct time functions, $y^1, y^2 \in Y$, there exists a bounded linear input-output system B^* which distinguishes these time functions; i.e., $B^*(y^1) \neq B^*(y^2)$.*

REFERENCES

1. D. Serakos, *Topics In Input-Output Systems Theory: Feedback Systems With Tapered Input Spaces, State and Generalized Adjoint Systems,* Ph.D. Dissertation, The University of Michigan, 1988.

2. W.L. Root and D. Serakos, "Generalized Adjoint Systems," in *Proceedings, 20th Princeton Conference on Information Sciences and Systems,* 1986.

3. W.L. Root, "Considerations Regarding Input and Output Spaces for Time-Varying Dynamical Systems," *Appl. Math. Optim.* **4** (1978), 365-384.

4. N. Dunford and J.T. Schwartz, *Linear Operators, Part I: General Theory,* Interscience Publishers, Inc., New York, 1957.

ON THE NEED FOR STABILIZING AND THE METHODS FOR STABILIZING REPETITIVE AND LEARNING CONTROLLERS

Richard W. Longman and Yi-Ping Hsin
Department of Mechanical Engineering
Columbia University, New York, New York 10027

"Linear Repetitive Controllers are Always Unstable". For repetitive control, a stability condition in the frequency domain is sometimes given that asks that the steady state frequency response of the z-transfer function from the error at one repetition to the error at the next repetition, have the property that the amplitudes of all frequencies decay every repetition [2] (for brevity, the reference numbers refer to those appearing in the citations in the single reference given below). In [7] it is shown that for repetitive control this is a sufficient condition for stability of the learning process, and if the number of time steps in a repetition is substantial, then it is very near to the stability boundary. In learning control it is often very far from the stability boundary, and in addition, when the trajectory is short compared to the settling time of the system, the derivation of the condition no longer applies. Nevertheless, in [2] it is seen that satisfying this condition is very desirable since it assures that one obtains monotonic convergence under appropriate conditions. Failure to satisfy this condition is very likely to lead to instability in repetitive control, and in learning control it is very likely to produce unacceptably bad transients. These transients often take the form of apparent initial convergence, followed by apparent divergence, followed by actual convergence [7]. The desirability of satisfying this condition both for repetitive and for learning control makes the distinction between these two types of control have relatively little practical importance for linear systems. The condition can be interpreted by requiring that the Nyquist plot of the digital repetitive/learning control law multiplied by the z-transfer function of the closed loop feedback control system, be inside the unit circle centered at +1 for all frequencies up to Nyquist. The radial distance from +1 to a point on the Nyquist plot represents how much the error component at that frequency is attenuated or amplified in one repetition. Very successful repetitive control laws have been designed and tested that use some form of phase compensation to keep the plot within the unit circle for a large frequency range, and then use some form of cutoff to attenuated signals associated with frequencies that fall outside the unit circle, and hence are associated with monotonic divergence ([4-6] and the reference below). We claim that such repetitive control laws are almost always unstable in the case of repetitive control, and subject to very bad transients in the case of learning control. There are several reasons for this. Usually, the phase cancellation is not perfect, making this cutoff necessary, and in addition the cutoff is not perfect, suggesting that small components outside the unit circle will make the repetitive control law unstable. In addition, in nearly every system, when one excites it at high frequencies, one can excite extra unmodeled dynamics, sometimes called parasitic poles. For example, one may use a rigid body model, but no body is truly rigid. Or one may model an amplifier as a pure gain, but at sufficiently high frequencies it does have dynamics and attenuation. These extra poles can produce extra multiples of 90 deg phase lag in continuous system models, and will easily send the Nyquist plot outside this unit circle centered at +1, no matter how well the phase of the original model is canceled. The instability can appear rather quickly as in Fig. 1, where the RMS tracking error decays by close to 3 orders of magnitude in a commercial robot performing a high speed maneuver, and then starts to diverge [3]. A third order model is used in this figure to perform the phase cancellation (the same model used in the numerical examples below). A 5th order model is determined from the data generated (there is a relationship between learning control and optimal experiment design, with learning control generating exactly the input needed to see the part of the system that is not being modeled well) and then learning is continued in Fig. 2. The error continues to decrease, but it is possible that one may need yet a better model later, including yet another vibration mode. In repetitive control the divergence can be very subtle, and require considerable patience in experimentation, as in the experiments in Fig. 3 run for 1000 repetitions without seeing an instability. Yet in Fig. 4 we run the experiment longer and start to see an instability around repetition 2,500 (see reference below).

Stabilization Methods for Repetitive Controllers. We conclude that there is a strong need for methods of stabilization for repetitive control, and for methods that produce good transients during learning in learning control. The authors and co-workers have suggested some. One is to do repeated identification as described for Figs. 1 and 2 [3]. A second method is to use a contraction mapping learning control law that has stability robustness for all sufficiently small learning gains (see Jang and Longman, *Advances in the Astronautical Sciences*, Vol. 93, Part 1, 1996). It can also be applied in repetitive control provided it is applied in a batch update mode with each update based on steady state response, but it loses its robustness when applied in real time repetitive control [4]. A third method is to produce stability by choice of finite word length as developed in the reference below. The fourth method is to apply zero-phase low-pass filtering on the accumulated learning signal in either learning or repetitive control [5]. In the latter case, one makes computations in real time, but the updates can only be applied when the computations are completed, i.e. in a batch update mode.

How Safe is Stabilization due to Quantization? The stabilization due to finite word length adjustment is very simple. The quantization level can be adjusted by use of frequency response information or it can be adjusted in hardware. In the reference below it is set so that steady state error of the feedback controller response for frequency components outside the unit circle are less than the quantization level, and hence are unable to accumulate. However, this procedure may be somewhat dangerous. Is it possible for nonsteady state situations to produce a divergent error? In Fig. 1, there is a local maximum produced in the error at one repetition of the learning process, that it attributed to a power serge that made all robot links run faster. Any frequency component of the response of the system to such a power serge that falls outside the unit circle is associated with monotonic growth of the associated steady state error each repetition. Hence, once some component of error at these frequencies occurs, that is larger than the quantization level, it will go into the accumulated signal, and it can never be removed (in the case of learning control it is removed when a transient wave of convergence arrives starting from the beginning of the trajectory and progressing to the end). Hence, there is a mechanism for accumulation of errors of a type that can never be eliminated.

The Stability Dangers. A single power surge probably has very small components in the frequency range outside the unit circle in any well design learning control law, so that although the error does not go away, it will remain small. Of course, it would be desirable to have a mechanism to eliminate the error even in this case. But the more important danger is for error histories that appear many times but are not periodic with the learning period. Examples include: 1) Start up. Suppose that the repetitive control law obtained at the end of the previous operation is applied when next turned on, after reaching steady state operation. Then a sudden introduction of the repetitive control signal produces something like a step change in the input and will produce something like a step response every time it is performed, i.e. every restart. If this error goes above the quantization level, then it will accumulate. 2) Each time a machine is turned on it warms up, and the error signature goes through a predictable time history, and components outside the unit circle larger than the quantization level (chosen based on steady state operating conditions) will accumulate. 3) Any repeating disturbance that has components at frequencies outside the unit circle that go above the chosen quantization level, and which are nonzero mean. 4) Wear on a machine may change the steady state error levels of the feedback control system producing more error at the frequencies outside the unit circle than the chosen quantization level will handle.

The Divergence Rate. The divergence rate of an unstable repetitive controller can be very slow, e.g. Fig. 4. In the reference, this divergence was eliminated by use of a sharper cutoff filter, which was then run for 10,000 repetitions without any sign of divergence. Theory would indicate that even this controller is unstable, but no instability is observed. One can show that the stability is due to the 7 digit word length used in the computer implementing the control action. In any machine that uses such a controller, and is operated hundreds of times a day, one would like to know whether the seemingly stable behavior will persist indefinitely, or whether one day one will come in and see that the machine is starting to diverge. The methods of the reference below can help assure one that the quantization level is correct for the steady state error of the feedback controller. But to be confident that there is no divergence, one needs to understand any repeating error signals that are not there every repetition as discussed above. If one is to establish this, not by analytical understanding as in the reference, but empirically, then one must be willing to test the process for many repetitions. Suppose that the source of error of concern is turning on the machine, and the step change in the input as the repetitive control signal is applied. If it takes 3,000 repetitions to see a significant accumulation of the associated signal, then one would have to have turned on the machine 3,000 times, which could represent a very much larger number of repetitions.

Solutions to Nonperiodic but Repeating Errors Above the Quantization Level. 1) Certain of the error sources listed above have specialized ways of avoiding difficulty. For example, to handle the step change in input produced when the previous repetitive control signal is first applied after start-up, one can apply this signal without using any repetitive control updates until after reaching steady state response. 2) If one sees divergence starting, one can always increase the quantization level until it stops the divergence. 3) One can use on-line computation of a zero-phase low-pass filter to repeatedly filter out the frequency components outside the unit circle with a batch update when the filtered result is ready [5]. This approach is the safest. It prevents divergence as long as each update attenuates these frequency components more than they accumulate between updates. Such filtering is also useful to eliminate any unwanted accumulated signal in 2.

Numerical Experiments. Here we numerically study how stabilization due to quantization behaves in several situations. Consider the third order model in the reference, with a constant unit input, disturbance sinusoids every 1 Hz up to 49 Hz, and quantization at 10^{-5}. First use the Case 2 repetitive control law of the reference, with a 20th order Butterworth low pass filter cutting off at 20 Hz, with linear phase lead and quantization. If we apply white Gaussian zero mean noise to the measurements, no accumulation is observed due to signals going above the quantization level. This might be expected since the noise is zero mean so that what is seen above the quantization level will average to zero. Digital hardware is often subject to occasional spurious measurements that are totally unrelated to the measured quantity. If we introduce such measurements randomly, having a uniform distribution between ±0.005 and a probability of occurrence at any one time step of 1/20,000, we observe no secular effect on the error after 20,000 repetitions of 100 time steps for a 1 sec repetition period. Again one might expect this based on the fact that the pulses have zero mean. However, we change the pulses to always be positive, uniformly distributed between zero and 0.015, and use a much simpler learning control law without high frequency filtering, consisting of a learning repetitive control gain of 0.1 and a linear phase lead of 5 time steps (Case 1 of the reference), causing the Nyquist plot to leave the unit circle around 20 Hz. The resulting RMS error for 20,000 repetitions is given in Figs. 5 and 6. Again no secular growth of error is observed. It is interesting to see the many different values of RMS error that become the steady state error between different pulses, and to note the relationship between the levels reached and the quantization level chosen. Now consider restarting the system every 400 repetitions. When restarting, the system is turned on, reaches steady state, and then the learning signal from the previous run is suddenly applied and the learning process started. The Case 2 model is used again, with the results shown in Figs. 7. No accumulation of error is observed. Finally, we simulate warm up. At repetitions 1000 and 2000 we simply change the system model by increasing the damping ratio to 1.0 to imitate increased friction when cold, and then decrease it in stages of 0.05 every 10 repetitions until reaching the original 0.5. Figure 8 gives the results. We see that instead of producing more error because of this repeating disturbance, we have less error. It appears similar to the changes in steady state error levels observed in Fig. 6. Note that in all of these cases, the frequency at which the system/repetitive controller leaves the unit circle is above the frequency of any dynamics in the system response.

We conclude that the finite word length repetitive control algorithm is more robust than might be expected. In order to illustrate possible difficulties, one must have the quantization level set to a tight limit on the steady state feedback control error. Then one must have a situation in which there are repeating errors that go above the quantization level, and do so in a systematic and nonzero mean manner, happening at the same point in the periodic cycle each time. And these must occur sufficiently often between times when one restarts the learning process, that significant accumulation of error is observed. Nevertheless, the zero-phase low-pass filtering method, either on-line batch or done before each start up is recommended.

Reference. Yi-Ping Hsin, R. W. Longman, E. J. Solcz, and J. deJong, "Stabilization due to Finite Word Length in Repetitive and Learning Control," AAS/AIAA Astrodynamics Conference, Sun Valley, Idaho, August 1997, to appear in *Advances in the Astronautical Sciences*, American Astronautical Society, Univelt Inc., Escondido, CA.

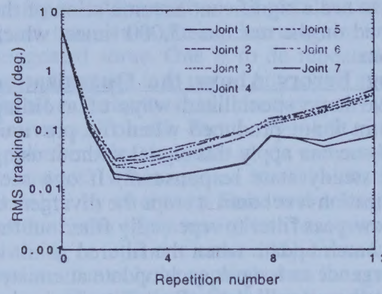


Fig. 1. RMS error, all joints, vs. repetitions.

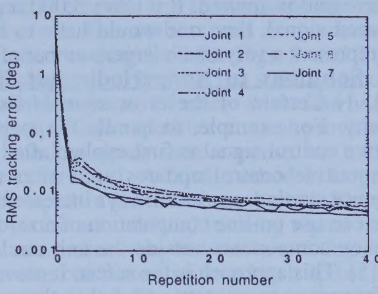


Fig. 2 Introduce one model update.

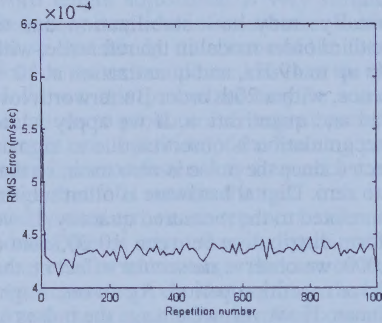


Fig. 3 An apparently stable system.

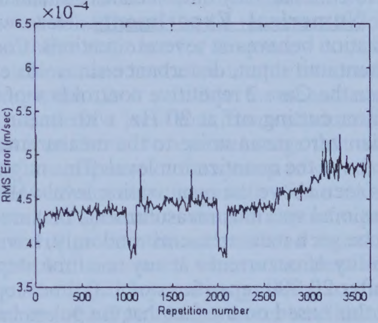


Fig. 4 Running longer shows it to be unstable

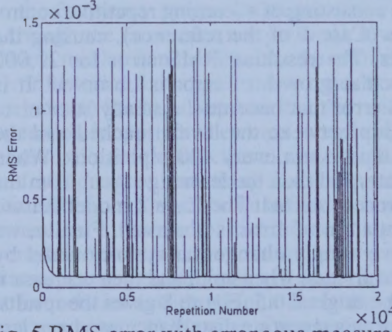


Fig. 5 RMS error with erroneous measurements.

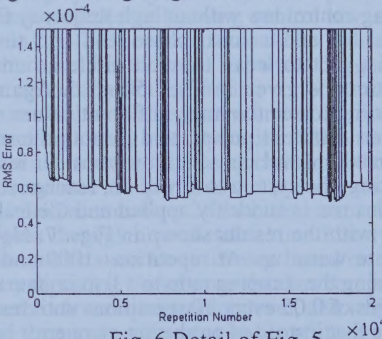


Fig. 6 Detail of Fig. 5.

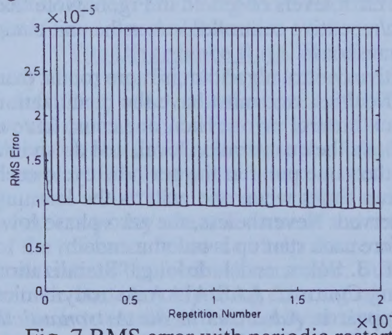


Fig. 7 RMS error with periodic restarts.

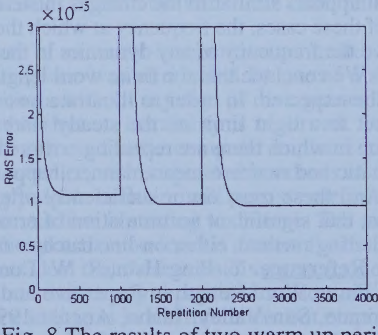


Fig. 8 The results of two warm up periods.

A Separation Property of Observer/Controller for Continuous-Time Fuzzy Systems

Ali Jadbabaie[*,1], Andre Titli[*,**], and Mo Jamshidi[*]
[*] *NASA Center for Autonomous Control Engineering (ACE)*
University of New Mexico, Albuquerque, NM 87131, USA

[**] *LAAS du CNRS and INSA Toulouse, France*

Email:{alij, titli,jamshidi}@unm.edu
http://ace.unm.edu

Abstract

In this paper, we present a systematic design methodology for design of observer based controllers for continuous-time fuzzy systems. We extend the previous results in this area which are based on state feedback to the case where states are not available for measurement, using an asymptotic observer. We show that the observer controller design can be done by solving a set of Linear Matrix Inequalities (LMIs), and the separation of the observer/controller holds.

Keywords: Takagi-Sugeno models, Linear Matrix Inequalities, fuzzy observers, separation property.

I. Introduction

Recently there has been a rapidly growing interest in using Takagi-Sugeno[1] fuzzy systems to approximate nonlinear systems. These type of systems differ from the so called "Mamdani" type fuzzy systems in that the consequents of Takagi-Sugeno (T-S for short, from now on) are analytic dynamical equations instead of fuzzy sets. The If-Then rules are obtained by linearization of nonlinear system around different operating points. Once the linear models are obtained, linear control methodology can be used to design local state feedback controllers for each rule. The global model is obtained by aggregation of these local fuzzy rules. The aggregated model can be viewed as a Polytopic Linear Differential Inclusion (PLDI) [2]. Sufficient conditions for stability of open loop T-S systems were first presented in [3]. These conditions dealt with the existence of a common positive definite matrix ***P*** which satisfy a set of Lyapunov Inequalities. Although finding a common positive definite matrix which would satisfy all these inequalities is by no means simple, This problem can be turned into an LMI feasibility problem[2]. Powerful interior point convex optimization methods exist for solving these LMIs [4]. Wang *et. al.* [5] used this approach to design stabilizing state feedback controllers. The problem of stabilizing T-S systems with a prescribed degree of stability was first addressed in [6]. Both of the above methods assume that all states are available for measurement. Unfortunately this is not the case in most practical applications. This may happen for several reasons. Measurement might be physically difficult, sensors are usually expensive, and are subject to failure and are noisy.

[1] Author of Correspondence

This motivates us to try to design out feedback controllers for these systems. To fulfill this task, we use an asymptotic observer for each local model. The overall observer would result from the aggregation of the T-S fuzzy rules. This observer, which is nonlinear in general (because of the non linearity of membership functions), can be designed separate from the controller using LMIs. In other words, the separation property holds for the observer/controller system.

The structure of this paper is as follows: First a couple of recent results in the stability analysis design of T-S fuzzy systems using Linear Matrix Inequalities (LMI's) are presented and their advantages and drawbacks are discussed. Next, the notion of *fuzzy observer* is introduced. Finally, it is shown that the *Separation Principle* holds for the continuous-time fuzzy observer/controller system.

II. Stability Analysis of Takagi-Sugeno Fuzzy Systems

This section deals with a brief review of continuous time T-S fuzzy systems.

Consider a nonlinear system modeled with a Takagi-Sugeno fuzzy model. This model is obtained by linearization of the nonlinear systems around different operating points. A generic rule can be written as follows:

i th Rule:

If x_1 is M_{i1} ... **and** x_n is M_{in} **THEN** $\dot{x} = A_i x + B_i u$

$$y = C_i x$$

$$i=1, \ldots, r,\ x \in R^n, u \in R^m, y \in R^p$$

where r is the number of rules, M_{ij} are the input fuzzy sets and $\begin{matrix} \dot{x} = A_i x + B_i u \\ y = C_i x \end{matrix}$ is the output of *i*th IF-THEN rule. Given a pair of input and state vectors and using the max-product inference method, the over all output of the fuzzy model can be written as follows:

$$\dot{x} = \frac{\sum_{i=1}^{r} \alpha_i(x)(A_i x + B_i u)}{\sum_{i=1}^{r} \alpha_i(x)} \tag{1}$$

Where $\alpha_i(x)$ is the overall truth value of the *i* th rule and is defined as follows:

$$\alpha_i(x) = \prod_{j=1}^{n} \mu_{ij}(x_j) \tag{2}$$

where μ_{ij} is the membership of x_j to the input fuzzy set M_{ij} The system (1) can be written as:

$$\dot{x} = \sum_{i=1}^{r} w_i(x)(A_i x + B_i u) \tag{3}$$

Where

$$w_i(x) = \frac{\alpha_i(x)}{\sum_{i=1}^{r} \alpha_i(x)} \tag{4}$$

$$\sum_{i=1}^{r} w_i(x) = 1$$

a sufficient stability condition for the open loop system (3) can be expressed by the following theorem :

Theorem 1 [3]: The equilibrium point of the fuzzy system (3) is globally asymptotically stable if there exist a common positive definite matrix ***P*** that satisfies the following inequalities:

$$A_i^T P + PA_i < 0 \qquad i = 1,2,\ldots,r \tag{5}$$

In order to analyze the stability of the closed loop system, a state feedback compensator is designed for each subsystem. The resulting fuzzy controller, which is generally nonlinear, is a fuzzy blending of each individual local controllers. For simplicity it is assumed that the fuzzy controller has the same input fuzzy sets as the system (1).

The fuzzy controller can be represented by a set of Takagi-Sugeno rules as follows:

Rule *i* : **If** x_1 is M_{i1} … **and** x_n is M_{in} **THEN** $\mathbf{u}(t) = -\mathbf{k}_i\, \mathbf{x}(t)$ (6)

Using the same argument, the over all controller can be written as:

$$u(t) = \frac{-\sum_{i=1}^{r} \alpha_i(x) k_i x}{\sum_{i=1}^{r} \alpha_i(x)} = -\sum_{i=1}^{r} w_i(x) k_i x \tag{7}$$

As it can be seen, the controller is nonlinear in general.

Substituting (7) in (1) we obtain

$$\dot{x} = \sum_{i=1}^{r} \sum_{j=1}^{r} w_i(x) w_j(x) (A_i - B_i k_j) x \tag{8}$$

Applying Theorem 1 to the closed loop system (8) results in the following theorem.

Theorem 2[5]: The equilibrium point of the closed loop system (8) is globally asymptotically stable if there exists a common positive definite matrix ***P*** and feedback gains k_i that the following conditions are satisfied:

$$(A_i - B_i k_i)^T P + P(A_i - B_i k_i) < 0 \qquad i<j \le r \quad ,j=1,\ldots,r \tag{9}$$

$$G_{ij}^T P + P G_{ij} < 0$$

$$Gij = \frac{(A_i - B_i k_j) + (A_j - B_j k_i)}{2}$$

The design problem is to find the controller gains k_i and a common positive definite matrix P such that the above conditions are satisfied. Multiplying both sides of the above inequalities by $P^{-1} = Q$, and introducing $Y_i = k_i Q$, the above inequalities can be recasted as the following LMIs:

$$\begin{aligned} & Q A_i^T + A_i Q - B_i Y_i - Y_i^T B_i^T < 0 \\ & Q(A_i + A_j)^T + (A_i + A_j)Q - B_i Y_j - Y_j^T B_i^T - B_j Y_i - Y_i^T B_j^T < 0 \end{aligned} \tag{10}$$

$$Y_i = k_i Q = k_i P^{-1}$$

$$j = 1, \ldots, r$$
$$i < j \le r$$

Wang *et al.* [5], have proposed a design methodology as follows:
First pick the local controller gains, then search for a common P using LMI's. If a common P can not be found change the controller gains. Repeat this process until a P can be found.

As it is evident, this approach is based on trial and error and it is highly probable that such a common matrix can not be found while the gains are fixed. Another approach proposed by Zhao *et al.* [6], is to find the controller gains and the common matrix P using the LMI method. Design objectives can be achieved by introducing a decay rate into equation (10) [6].

III. Fuzzy Observers

In the previous section, we discussed existing methods for designing stabilizing state feedback controllers for T-S fuzzy systems , however, in many practical applications, measuring the states is very difficult and costly, and obtaining the Takagi-Sugeno rules from state measurements is not always simple. In addition, the number of rules grow exponentially with the number of states.

In order to fix this problem a set of Takagi-Sugeno rules can be used to estimate the states. The notion of *fuzzy observer* is introduced.

Definition : A fuzzy observer is a set of T-S rules that can estimate the states of the local subsystems using an asymptotic observer methodology. A generic rule can be expressed as

Rule i: **If** $y_1(t)$ is M_{i1} … and $y_p(t)$ is M_{ip} **THEN:** $\dot{\hat{x}} = A_i\hat{x} + B_i + L_i(y-\hat{y})$

$$\hat{y}_i = C_i\hat{x}$$

where L_i*'s* are the local observer gains, and the pairs (A_i, C_j) are observable.

$$y = \sum_{i=1}^{r}\alpha_i(y)C_i x(t)$$

is the measured output, and

$$\hat{y} = \sum_{i=1}^{r}\alpha_i C_i\hat{x}$$

is the estimated output.

The overall observer can be written as :

$$\dot{\hat{x}} = \sum_{i=1}^{r}\sum_{i=1}^{r}\alpha_j(x)\alpha_i(x)(A_i\hat{x} + B_i u - L_i C_j(x-\hat{x}))$$

Since:

$$u(t) = -k_i\hat{x}(t)$$

substituting for $u(t)$, equation (3) becomes the following:

$$\dot{x} = \sum_{i=1}^{r}\alpha_i(x)A_i x - \sum_{i=1}^{r}\sum_{j=1}^{r}\alpha_i(x)\alpha_j(x)B_i k_j\hat{x} \tag{11}$$

By introducing the estimation error as a new variable, the following equations are obtained:

$$\tilde{x} = x - \hat{x} \tag{12}$$

$$\dot{\tilde{x}} = \sum_{i=1}^{r}\sum_{j=1}^{r}\alpha_i(x)\alpha_j(x)(A_i - L_i C_j)\tilde{x}(k)$$

The observer dynamics is stable if a common positive definite matrix P_2 is found that satisfies the following inequalities:

$$(A_i - L_i C_i)^T P_2 + P_2(A_i - L_i C_i) < 0 \qquad i<j \le r,\ j=1,\dots,r \tag{13}$$

$$H_{ij}^T P_2 + P_2 H_{ij} < 0$$

$$H_{ij} = \frac{(A_i - L_i C_j) + (A_j - L_j C_i)}{2}$$

Obviously the above matrix inequalities are not LMI's in P_2 and L_i's, but with a simple change of variables we can turn the above inequalities into LMI's.

Let $Z_i = P_2 L_i$

The conditions in (13) can be written as the following LMIs in P_2 and Z_i

$$A_i^T P_2 + P_2 A_i - Z_i C_i - C_i^T Z_i^T < 0 \tag{14}$$

$$(A_i + A_j)^T P_2 + P(A_i + A_j) - (Z_i C_j + C_j^T Z_i^T) - (Z_j C_i + C_i^T Z_j^T) < 0$$

$$Z_i = P_2 L_i$$

Theorem 3: Let Σ be the closed-loop system defined as follows:

Σ:

$$\begin{bmatrix} \dot{x} \\ \dot{\tilde{x}} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{r} \sum_{j=1}^{r} w_i w_j (A_i - B_i k_j) & \sum_{i=1}^{r} \sum_{j=1}^{r} w_i w_j B_i k_j \\ 0 & \sum_{i=1}^{r} \sum_{j=1}^{r} w_i w_j (A_i - L_i C_j) \end{bmatrix} \begin{bmatrix} x(t) \\ \tilde{x}(t) \end{bmatrix} \tag{15}$$

The closed-loop fuzzy system (15) is globally asymptotically stable if a common positive definite matrix $\tilde{\mathbf{P}}$ and controller and observer gains k_i and L_i can be found that satisfy the following Matrix Inequalities :

$$\tilde{A}_{ii}{}^T \tilde{P} + \tilde{P} \tilde{A}_{ii} < 0 \qquad i < j \le r \qquad j = 1, \ldots, r \tag{16}$$

$$\left(\frac{\tilde{A}_{ij} + \tilde{A}_{ji}}{2}\right)^T \tilde{P} + \tilde{P}\left(\frac{\tilde{A}_{ij} + \tilde{A}_{ji}}{2}\right) < 0$$

Where $\tilde{A}_{ij}$ is defined as follows: $\tilde{A}_{ij} = \begin{bmatrix} A_i - B_i k_j & B_i k_j \\ 0 & A_i - L_i C_j \end{bmatrix}$

The proof directly follows from Theorem 2. Note that the above conditions are not LMI's, but can be transformed into LMI's using the same change of variables as in equations (10) and (14). Moreover, it would be very important to design for observer and controller gains independently. In other words, it is desirable to find a block diagonal $\tilde{P}$ that satisfies the conditions in (16). Therefore, we would like to show that conditions in (16) are equivalent to conditions in (9) and (13). This is shown in the following main result:

Theorem 4 (Main Result: separation Theorem for T-S fuzzy systems): The closed loop system (15) is globally asymptotically stable if the conditions in (9) and (13) are satisfied.

Proof: Let us choose $\tilde{P}$ as a block diagonal matrix with λQ^{-1} and P_2 as block matrices on the diagonal, where λ is a positive constant.

$\widetilde{P} = \begin{bmatrix} \lambda Q^{-1} & 0 \\ 0 & P_2 \end{bmatrix}$. We will show that there always exists a $\lambda > 0$ such that $\widetilde{P}$ satisfies conditions in (16), providing the conditions (9) and (13) are satisfied. Using the above $\widetilde{P}$ in the Lyapunov Inequality, we have

$$\widetilde{A}_{ii}{}^{T}\widetilde{P} + \widetilde{P}\widetilde{A}_{ii} < 0$$

substituting for $\widetilde{P}$ and $\widetilde{A}_{ii}$, we obtain the following matrix inequality :

$$\begin{bmatrix} \lambda\left[(A_i - B_i k_i)^T Q^{-1} + Q^{-1}(A_i - B_i k_i)\right] & \lambda Q^{-1}(B_i k_i) \\ \lambda (B_i k_i)^T Q^{-1} & (A_i - L_i C_i)^T P_2 + P_2 (A_i - L_i C_i) \end{bmatrix} < 0 \tag{17}$$

Using the LMI lemma [7], the above matrix is negative definite if and only if the following conditions are satisfied:

$$\lambda\left[(A_i - B_i k_i)^T Q^{-1} + Q^{-1}(A_i - B_i k_i)\right] < 0 \tag{18}$$

$$\lambda Q^{-1}(B_i k_i)[(A_i - L_i C_i)^T P_2 + P_2(A_i - L_i C_i)]^{-1}(B_i k_i)^T Q^{-1} - \left[(A_i - B_i k_i)^T Q^{-1} + Q^{-1}(A_i - B_i k_i)\right] > 0$$

Since $\left[(A_i - B_i k_i)^T Q^{-1} + Q^{-1}(A_i - B_i k_i)\right] < 0$, the first condition is already satisfied. Moreover, we can always pick λ small enough such that the second condition is also true.
(Note that: $-\left[(A_i + B_i k_i)^T Q^{-1} + Q^{-1}(A_i + B_i k_i)\right] > 0$)

Since both conditions of LMI lemma are satisfied, the matrix in (17) is negative definite.
Using the same argument, it can be shown that $\widetilde{P}$ satisfies the second condition in (16) as well, that is, $\left(\frac{\widetilde{A}_{ij} + \widetilde{A}_{ji}}{2}\right)^T \widetilde{P} + \widetilde{P}\left(\frac{\widetilde{A}_{ij} + \widetilde{A}_{ji}}{2}\right) < 0$

As a result, the closed loop system is asymptotically stable, and the proof is complete.

IV. Conclusion

The purpose of this paper was to extend some recent results on the stability of Takagi Sugeno fuzzy systems to the case were states are not available for measurement. We showed that the fuzzy observer can be obtained by aggregation of local observers for each linear model. Moreover, we presented a separation theorem, and proved that the observer and controller can be designed separately.

V. Acknowledgment

The authors would like to thank Dr. Peter Dorato for his helpful suggestions.

References

[1] T. Takagi and M. Sugeno, " Fuzzy Identification of systems and its application to modeling and control," *IEEE Trans. Syst.,Man, Cybern.*, vol. 15 , pp-116-132, 1985.

[2] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in systems and control theory*, SIAM series in Applied mathematics, Philadelphia, PA, *1994*

[3] K. Tanaka and M. Sugeno, " Stability analysis of fuzzy systems using Lyapunov's direct method," *Proc.* NAFIPS '90,1990, pp. 133-136.

[4] Y. Nesterov and A. Nemirovsky, *Interior point polynomial methods in convex programming, SIAM* , Philadelphia 1994.

[5] H. Wang, K. Tanaka and M. Griffin, " An spproach to fuzzy control of nonlinear systems: Stability and design issues, " *IEEE Trans. Fuzz. Syst.*, vol.4 , No.1, pp. 14-23, 1996.

[6] J. Zhao, V. Wertz and R. Gorez, " Fuzzy gain scheduling controllers based on fuzzy models," Proc. *Fuzz IEEE ' 96*, New Orleans, LA, 1996, vol. 1, pp- 16700-1676.

[7] P. Dorato, C. T. Abdallah, and V. Cerone, *Linear Quadratic Control : An Introduction*, Prentice Hall , Englewood Cliffs, NJ, 1995.

On the Local Polynomial Approximation of the Normal Form: The SISO Case.

Jesús A. Esquivel
FIME UAdeC
Barranquilla s/n. Col. Guadalupe
Monclova, Coahuila. c.p. 25750
MEXICO
jesquive@alpha3.mon.uadec.mx

Joaquín Collado
FIME UANL
Apdo. Postal F-30, c.p. 66450
San Nicolás de los Garza. N.L.
MEXICO
jcollado@ccr.dsi.uanl.mx

1 Introduction

In nonlinear control with differential geometry approach, to get a canonical normal form is a most for many problems. From the computational point of view, this represents two main difficulties. The first is to get the inverse of a given nonsingular transformation (diffeomorphism), an approximate polynomial solution was proposed by the authors [1] to overcome this problem. The second, that the aforementioned transformation satisfies an extra condition, namely $L_g\phi_i(x) \equiv 0$, in order to get the normal form. This condition is equivalent to the solution of a system of Ordinary Differential Equations as advanced in [2], [6] and [5]. We have developed a procedure based in a lemma presented in [6] where a change of coordinates can be used to see that the selection of the functions is immediate. As a polynomial approximation to the solution of an ordinary differential equation around of a given point, belongs to a class defined by this integral curve (see [3]), this means that the polynomial approximation can be used indifferently in the mappings involved in the change of coordinates as it were the true solution in the given chart.

2 Preliminaries

We consider the *local analysis* of single input/single output systems of the form

$$\begin{aligned} \dot{x} &= f(x) + g(x)u \\ y &= h(x) \end{aligned} \tag{1}$$

where $x \in R^n$, u, $y \in R$, f and g are smooth vector fields on R^n and $h : R^n \to R$ is a smooth function. We have based on the Lemma 1.2 and the Proposition 1.4 given in the Chapter 4 of [2]

Lemma 1 *[6]Suppose $\triangle(x) = span\{f(x)\}$, $x_0 \in X$ (open subset of R^n) and $f(x_0) \neq 0$. Then there exist a neighborhood $U \subseteq X$ of x_0 and a diffeomorphism $T : U \to X$ such that*

$$f_T(y) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}', \forall y \in T(U)$$

We can apply this lemma to see that is *necessary* to choose the functions independent of the first new variable, y_1, to satisfy the condition, so it is sufficient to take the functions as $\phi_i(y) = y_i$, $i = 2, \ldots, n$, hence in the original coordinates would be $\phi_i \circ T(x) = \phi_i \circ q^{-1}(x) = y_i = T_i(x)$, $i = 2, \ldots, n$ and in the proof of the Proposition 1.4 of [2] is demonstrated that $(n - r)$ of these $(n - 1)$ functions become the last part of the diffeomorphism.

3 Example

Consider the system $f(x) = [x_2, 0, 0]'$, $g(x) = [x_1 - x_3, 1 + \cos(x_2), -\exp(x_1) * (1 + x_3)]'$, $y = x_1$ with $x_0 = [0, 0, 0]'$. The relative degree of this system is two and the first two terms of the diffeomorphism to obtain the normal form are $\phi_1(x) = h(x) = x_1$ and $\phi_2(x) = L_f h(x) = x_2$. To complete the diffeomorphism is necessary to find $\phi_3(x)$ such that $L_g \phi_3(x) = 0$, and because of the difficulty to get it, we would wish at least a local approximation. The transformation becomes, keeping terms up to second degree.

$$\begin{aligned}\Phi(x) = [&\mathit{x1}, \mathit{x2}, .01410709442\,\mathit{x2} + .00001015780401\,\mathit{x1} + .1506595608\,\mathit{x3} \\ &+ .0001463011203\,\mathit{x1}^2 + .04455518772\,\mathit{x1}\;\mathit{x2} + .004600535843\,\mathit{x1}\;\mathit{x3} \\ &+ .03707774382\,\mathit{x2}^2 + .01817840352\,\mathit{x2}\;\mathit{x3} + .01190692779\,\mathit{x3}^2]\end{aligned}$$

which is used to approximate locally to the normal form.

4 Conclusions

The method proposed has been proved in several nontrivial examples and the results show that may be used to explore, at least locally, the properties of the normal form and with the diffeomorphism involved, to analyze the corresponding properties in the original space.

References

[1] Collado, J. and J.A. Esquivel. *On the Symbolic computation in Nonlinear Systems.* Accepted in a IFAC System and Structure Conference. Bucharest, Rumania. Oct. 1997.

[2] Isidori, A. . *Nonlinear Control Systems: An Introduction.* Springer Verlag. 3rd Edition. 1996.

[3] Jurdjevic, V. . *Geometric Control Theory.* Cambridge Studies in Advanced Mathematics 51. 1997.

[4] Nijmeijer, H. and A.J. Van der Schaft. *Nonlinear Dynamic Control Systems.* Springer Verlag New York . 1990.

[5] Rantzer, A. . Nonlinear Control Lecture Notes. Lund Institute of Technology. 1995.

[6] Vidyasagar, M. . *Nonlinear Systems Analysis.* Prentice Hall, New Jersey. 2nd Edition. 1993.

LIST OF AUTHORS

Aazhang, B. 460
Akl, S.G. 585
Alexander, P.D. 187
Amato, F. 643,656,888
Amir, E. 759
Anantharam, V. 233
Aracil, J. 261
Asenstorfer, J.A. 187
Ashley Jr., J. 513
Balakrishnan, H. 759
Bao, J.Q. 64
Barbarossa, S. 385
Barbeau, M. 533
Barg, A. 681
Baum, K.L. 719
Bertsekas, D.P. 913
Biglieri, E. 1023
Blake I.F. 196
Bock, H.G. 817,845
Bouthellier, P.R. 623
Brandt-Pearce, M. 321
Braslavsky, J.H. 858
Brown, D.R. 54
Caire, G. 1023
Calabrò, S. 173
Calderbank, A.R. 440
Campanella, M. 173
Casas, R.A. 769
Celentano, G. 656
Chang, J.-H. 93
Chen, H. 147
Chen, T.-T. 311
Chu, L.-C. 330
Coffey, J.T. 147
Collado, J. 1079
Coraluppi, S.P. 943
Costello, Jr., D.J. 711
Davison, D.E. 848
De Leòn, P.L. 122
Deller, Jr., J.R. 789
Dumer, I. 976
Duursma, I.M. 41
Edell, R. 261
Elkies, N.D. 23
Elliott, R. 779
Elmasry, G.F. 126
Ephremides, A. 583
Ericson, T. 413
Esquivel, J.A. 1079
Fabian, M. 543
Feldmann, A. 269
Feng, G.-L. 205
Fevrier, I.J. 1003
Fields, J. 422
Fischer, R. 966
Fitz, M.P. 1003
Ford, J. 779
Frappier, M. 533
Freudenberg, J.S. 646,868
Frey, B.J. 666,691
Fuhrmann, D.R. 291
Fujii, T. 835
Fujiwara, T. 185
Ganti, A. 739
Garbo, G. 173
Gelfand, S.B. 807,1003
Ghosh, B.K. 623
Giannakis, G.B. 385,729
Gibson, J.D. 128
Gilbert, A.C. 269
Goeckel, D.L. 370,383
Goodwin, G.C. 898
Green, M. 769
Gu, K. 631
Hafeez, A. 493
Hanly, S. 281
Hartel, B. 817
Helleseth, T. 395
Hochwald, B.M. 1013,1033
Høholdt, T. 229
Holloway, L.E. 513
Honig, M.L. 83
Hopp, W.J. 553
Hsin, Y.-P. 1067
Huber, J. 966
Huffman, C.W. 404
Hughes, B.L. 360
Jacobs, S.P. 112
Jacobson, C.A. 769
Jacquet, P. 242
Jadbabaie, A. 1071
Jamshidi, M. 1071
Jelenković, P.R. 1013
Ji, C. 138
Joachim, D. 789
Johnson, Jr., C.R. 54,769
Justesen, J. 961
Kabamba, P.T. 848
Kabanza, F. 533
Kanai, K. 827
Katz, R.H. 759
Ke, N.-P. 878
Kedia, V. 112
Koetter, R. 44
Kokotović, P.V. 858
Kong, X. 798
Krasikov, I. 156
Krogmeier, J.V. 807
Kschischang, F.R. 666
Kumar, P.V. 395
Kumar, R. 543
Kuo, S.M. 798
Kuroda, I. 993
Kurtz, T.G. 269
Kwon, S. 291
LeBlanc, J.P. 122
Lehnert, J.S. 311
Li, J. 845
Li, W.-C.W. 14
Liberty, S.R. 933
Litsyn, S. 156
Loeliger, H.-A. 666,687
Longman, R.W. 817,845,1067
Luo, W. 583
Lynch, R. 124
Ma, S. 138

MacKay, D.J.C. 691
Malesinska, E. 573
Malhamé, R.P. 523
Marcus, S.I. 943
Marzetta, T.L. 1033
Mattei, M. 888
Mayne, D.Q. 858
McKeown, N. 595
Médard, M. 1043
Meerkov, S.M. 848
Megalooikonomou, V. 102
Middleton, R.H. 868
Min, P.S. 563
Mitra, U. 330
Mohseni, J. 923
Moore, J.B. 779
Moreno, C.J. 1
Naguib, A. 440
Nakashima, K. 835
Nassar, C.R. 340
Nayeri, M. 789
Nishitani, T. 993
O'Sullivan, J.A. 112
Öberg, M. 701
Oh, M.-S. 563
Oh, S.-J. 380
Olejniczak, K.J. 923
Ortiz, D.S. 646
Osa, Y. 827
Osadciw, L.A. 350
Padmanabhan, V.N. 759
Panwar, S.S. 604
Papadopoulos, H.C. 483
Parhi, K.K. 993
Pellikaan, R. 33
Pironti, A. 888
Pless, V. 422
Poor, H.V. 450
Pottie, G.J. 74
Prabhakar, B. 595
Pujara, L.R. 614
Pursley, M.B. 713
Quast, P. 933
Rao, R.R. 749
Rao, T.R.N. 205
Rashid-Farrokhi, F. 93
Ratasuk, R. 83
Reed, M.C. 187
Rey, G.J. 769
Rong, C. 395
Rosenthal, J. 953
Roy, R. 604
Ryu, B. 252
Sahr, J.D. 301
Sain, M.K. 933,1053
Sain, P.M. 1053
Sakata, S. 231
Saquib, M. 739
Scaglione, A. 385,729
Schniter, P.B. 54
Schultz, V.H. 845
Sen, A. 573
Sendonaris, A. 460
Senturk-Gel, E. 553
Serakos, D. 1063
Seron, M.M. 858
Seshadri, N. 440
Seshan, S. 759
Shamai, S. 470
Shi, X. 205
Shokrollahi, M.A. 225
Siegel, P.H. 701
Sinha, A.S.C. 641
Smarandache, R. 953
Soleymani, M.R. 340
Song, L. 993
Spencer, Jr., B.F. 1053
St-Denis, R. 533
Stark, W.E. 493
Steinbach, M. 817
Stemm, M. 759
Subbarao, M.W. 360
Sudan, M. 215
Takeshita, O.Y. 711
Tanaka, K. 827
Taricco, G. 1023
Tarokh, V. 440
Tassiulas, L. 93
Thistle, J.G. 523
Titlebaum, E.L. 350
Titli, A. 1071
Tong, L. 64
Tse, D. 281
Tseng, D.-F. 503
Tsujino, T. 835
Uchida, T. 185
Uchikado, S. 827
Uppala, S.V. 301
Van Oyen, M.P. 553
Varaiya, P. 261
Verdú, S. 83,470
Wachsmann, U. 966
Wang, X. 450
Wasserman, H. 225
Wasserman, K.M. 380
Wei, Y. 807
Weiner, P.A. 183
Wiberg, N. 666
Wilkins, C.S. 713
Willett, P. 124
Willinger, W. 269
Willson, Jr., A. 986
Wolff, T. 585
Wood, J.A. 163
Woodyatt, A.R. 868
Wornell, G.W. 483
Yamamoto, H. 185
Yang, K. 395
Yang, M.-H. 321
Yang, W. 128
Yates, R. 739
Yaz, E.E. 923
Yesha, Y. 102
Ytrehus, Ø. 432
Zhou, S. 681
Zhu, J. 205
Zhu, M. 595
Zoltowski, M.D. 503
Zorzi, M. 749

Mahaffy, D.S.C. 881
Makarenko, R. 774
Marchetti, S.P. 223
Martin, M. 907
Martin, J.A. 1033
Mason, M. 576
Massie, D.S. 951
McKenney, R. 795
Mehrez, M. 1043
Menshikov, S.M. 608
Meyer-Hermann, V. 912
Miyamoto, K. 818
Mok, P.S. 761
Moon, H. 500
Morrison, J. 923
Muller, S.B. 776
Murata, T. 13
Nagata, A. 666
Nagashima, K. 837
Nicole, C.R. 159
Nguyen, M. 709
Nishioka, T. 967
O'Flaherty, T.A. 413
Ohara, M. 1003
Oh, S. 565
Oh, E.J. 380
Okkawa, K.A. 772
Ono, D.A. 548
Ozer, V. 927
Ozdemir, E.A. 980
Papadopoulos, H. 779
Pasini, S. 989
Papageorgiou, D.E. 441
Park, R.K. 671
Pelinovsky, E. 24
Petrov, A. 788
Pires, V. 422
Platt, H.V. 830
Potter, G.S. 15
Prokhorov, B. 1093
Purcell, L.R. 539
Quinn, M.H. 753
Quin, F. 992
Rao, K.R. 748
Rao, T.S.R. 809
Raspail-Fendel, E. .....
Roberts, R. 53
Reed, M.C. 687
Riley, D.J. 765
Rong, C. 195
Rosenthal, J. 1053
Ryu, K. 104
Ryu, B. 232
Sabe, A.D. 71
Sahu, M.K. 1033
Sain, P.M. 1031
Sasaki, S. 731
Sayari, M. 739
Scaglione, A. 523
Schreier, P.B. 54
Schulz, V.F. 168
Sen, A. 575
Shahbazian, A. 455
Simmons-Clark, R. 231
Sinclair, A. 1093
Sheng, M.M. 558
Sermon, P. 440
Shapiro, S. 728
Shibata, K. 418
Shi, X. 509
Shekhtman, M.A. 222
Singer, P.H. 701
Sinha, A.S.C. 247
Sobolev, M.V. 957
Soloyanov, M.V. 190
Song, Y. 883
Spencer, H., B.F. 1058
St Pierre, R. 317
Stern, W.J. 495
Strohmeyer, M. 517
Suzuki, M. 139
Sukhanov, M.V. 150
Suzuki, M. 275
Takahashi, H.F. 715
Tanaka, K. 819
Tayama, G. 913
Teller, P. 891
Thomas, C. 15
Thurber, J.D. 887
Tikhonov, B.J. 990
Toth, B. 1073
Tsang, L. 303
Tsu, D. 217
Turner, S.F. 553
Ueda, F. 817
Ulrich, T. 129
Urbanek, S. 929
Utkin, A.V. 951
Van Dyke, M.D. 713
Vasquez, P. 263
Verma, A. 470
Vinogradov, P. 566
Wang, A. 880
Wasserman, H. 273
Wasserman, T.M. 585
Wei, F. 597
Weiss, F.S. 149
White, B. 509
Williams, C.S. 711
Willis, F. 132
Wilmington, W. 718
Wilson, T.J. 964
Wolf, T. 113
Wong, J.A. 163
Woolgar, A.S. 963
Wright, G.W. 823
Yamazaki, H. 182
Yang, F. 293
Yang, M.-H. 421
Yang, W. 123
Yates, R. 795
Ye, T.J. 719
Yoshida, Y. 109
Yurtsel, G. 774
Zhou, S. 831
Zhu, L. 305
Zhu, M. 587
Zolotarev, M.D. 568
Zoom, M. 749

UNIVERSITY OF ILLINOIS-URBANA

3 0112 033279784